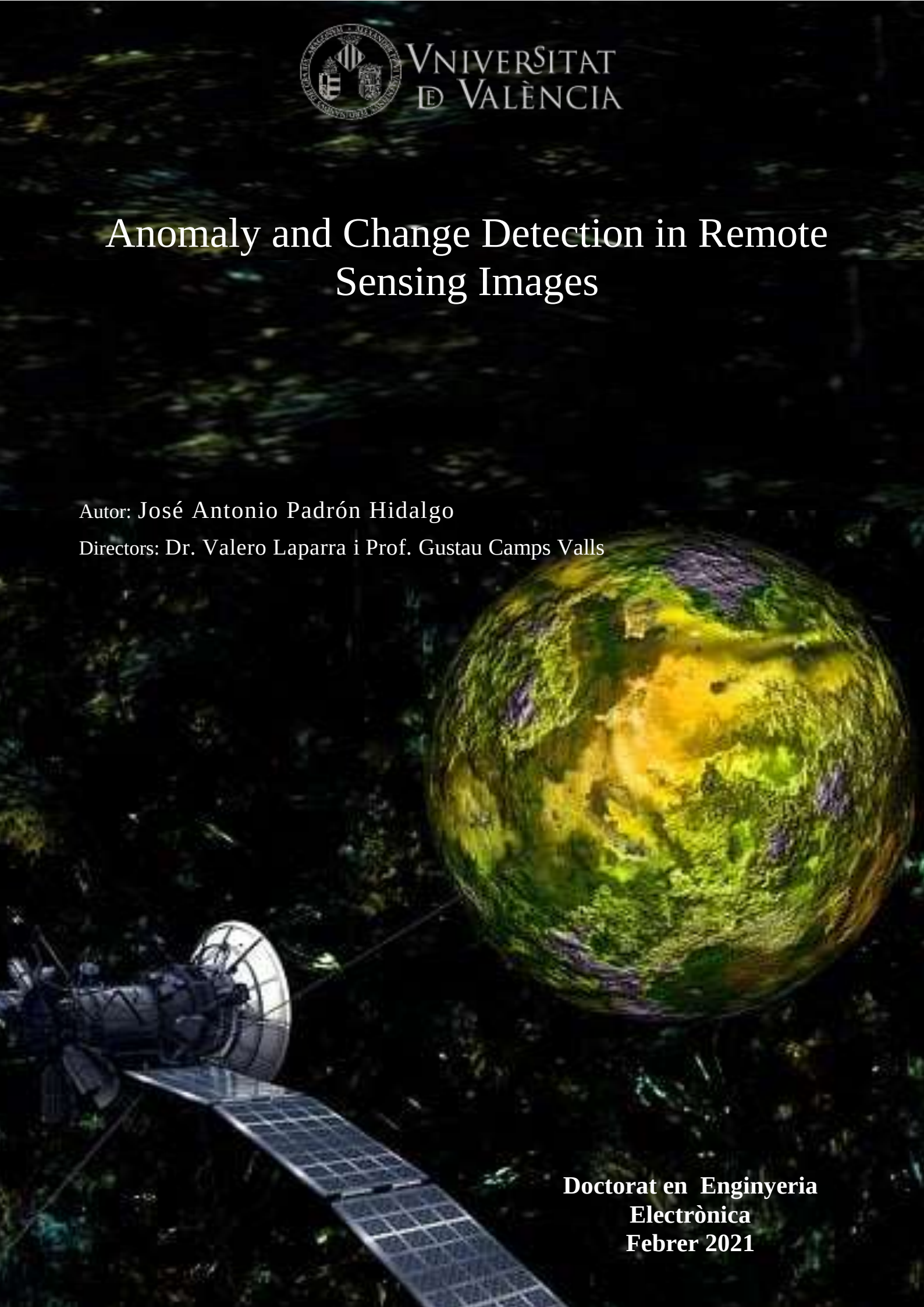


Anomaly and Change Detection in Remote Sensing Images

Autor: José Antonio Padrón Hidalgo

Directors: Dr. Valero Laparra i Prof. Gustau Camps Valls

Doctorat en Enginyeria Electrònica
Febrer 2021

TESI DOCTORAL EN ENGINYERIA ELECTRÒNICA

ESCOLA TÈCNICA SUPERIOR D'ENGINYERIA

ANOMALY AND CHANGE DETECTION IN REMOTE SENSING IMAGES

PER

JOSÉ ANTONIO PADRÓN HIDALGO

Directors:
DR. VALERO LAPARRA
PROF. GUSTAU CAMPS VALLS

Doctorat en Enginyeria Electrònica
Universitat de València
Febrer-2021

DR. VALERO LAPARRA, PhD in Computer Science and Computational Mathematics, Assistant professor (Professor Ajudant Doctor) at the Departament d'Enginyeria Electrònica in Escola Tècnica Superior D'Enginyeria at the Universitat de València

PROF. GUSTAU CAMPS VALLS, PhD in Physics, Full professor (Catedràtic d'Universitat) at the Departament d'Enginyeria Electrònica in Escola Tècnica Superior D'Enginyeria at the Universitat de València

DECLARE THAT:

The Telecommunications and Electronics engineer José Antonio Padrón Hidalgo has developed under his direction the work entitled *Anomaly and Change Detection in Remote Sensing Images*, which is presented in this document to qualify for the PhD degree from the University of Valencia.

And so that it may be recorded for the appropriate purposes, and giving the approval for the presentation of this work before the corresponding Thesis Tribunal, we sign this certificate in Valencia on July 2, 2021.

Valero Laparra

Gustau Camps Valls

DOCTORAL THESIS:
Anomaly and Change Detection in Remote Sensing Images

AUTHOR:
José Antonio Padrón Hidalgo

DIRECTORS:
Dr. Valero Laparra
Prof. Gustau Camps Valls

El tribunal nombrat per jutjar la Tesi Doctoral citada anteriorment, compost per:

President: _____

Vocal: _____

Secretari: _____

Acorda otorgar-li la qualificació de _____

I per a què així conste a efectes oportuns, signem el present certificat.

A Burjassot, el de de 2021

NOTE TO THE READER

According to the University of Valencia Doctorate Regulation [1] this PhD dissertation is presented as a compendium of at least three publications in international journals containing the results of the conducted work. It also describes work that has recently been submitted to scientific journals. Furthermore, in accordance with the aforementioned regulation and with the aim to foster the language of the University of Valencia in research and education activity, this PhD dissertation starts with two abstracts in Spanish and Valencian. In addition, an summary dissertation in Spanish is included at the end of the Thesis.

[1] Reglament sobre depòsit, avaluació i defensa de la tesi doctoral aprovat pel Consell de Govern de 28 de Juny de 2016. ACGUV 172/2016.
Pla d'increment de la docència en valencià (ACGUV 129/2012) aprovat i modificat pel Consell de Govern de 22 de desembre de 2016. ACGUV 308/2016.

Agradecimientos

En primer lugar, me gustaría agradecer a mi director de Tesis Gustau Camps-Valls por darme la oportunidad de formar parte de su grandioso equipo de trabajo. Quien con sus conocimientos y apoyo me guió a través de cada una de las etapas de este proyecto para alcanzar los resultados que buscaba. A Valero Laparra por su ayuda incondicional desde el primero momento que lo necesité. Además de tutor, compañero inseparable que me guió y me ayudó a encontrar los buenos caminos y me proporcionó todas las herramientas necesarias para llevar esta Tesis al lugar que corresponde.

También quiero agradecer a la fundación Santiago Grisolia de la Generalitat Valenciana que hizo posible el desarrollo de este proyecto en colaboración con el grupo de procesamiento de imágenes (IPL) de la Universitat de València. Quiero agradecer por brindarme todos los recursos y herramientas que fueron necesarios para llevar a cabo el proceso de investigación. No hubiese podido arribar a estos resultados de no haber sido por su incondicional ayuda.

Tengo la suerte de tener un excelente grupo de trabajo en el cual encontré una familia. Agradecer a todos mis amigos estudiantes que al igual que yo, estamos en el proceso de obtener un grado de Doctor. No olvidaré ningún momento o experiencia vivida a sus lados. Recordaré nuestros coffe time, nuestras fiestas, así como los momentos difíciles por los que pasamos. También quiero agradecer en especial a Fatih Nar y Adrián Pérez-Suay por aceptar ser parte de este proyecto y por todas las enseñanzas que me transmitieron en todo el proceso.

Por último, pero no menos importante quiero agradecer a mi familia, por apoyarme aun cuando mis ánimos decaían. En especial, quiero hacer mención a mi madre Elaine María y con un astérisco a mi hermana Lisandra Muñoz, que siempre estuvieron ahí para darme palabras de apoyo y energías. Les estaré infinitamente agradecidos y quiero que sepan que son de las personas más importantes de mi vida. Además, quiero agradecer el apoyo incondicional de mi prometida Ana Lucía quien a pesar de la distancia siempre confió en mí y me apoyó hasta el final de esta larga carrera. A mis amigos cubanos que estuvieron a mi lado todos estos años a pesar de las adversidades y de estar lejos del lugar que nos vio nacer. Muchas gracias a todos.

Contents

List of Figures

List of Tables

Acronyms

ACD	Anomalous Change Detection
AD	Anomaly Detection
AUC	Area Under Curve
AVHRR	Advanced Very High Resolution Radiometer
AVIRIS	Airborne Visible/Infrared Imaging Spectrometer
CD	Change Detection
CVA	Change Vector Analysis
DRCOG	Denver Regional Council of Government
EC	Elliptically Contoured
EO	Earth observation
EUMETSAT	European Organization for the Exploitation of Meteorological Satellites
$\mathbb{H}$	Hilbert Space
GSD	Ground Sample Distances
HR	High Resolution
K	Kernel
KC	Kernel Cook
KM	Kernel Machine
KRR	Kernel Ridge Regression
KRX	Kernel RX
KSC	Kennedy Space Center

LASSO	Least Absolute Shrinkage and Selection Operator
LDCM	Landsat Data Continuity Mission
MAD	Multivariate Alteration Detection
MODIS	Moderate Resolution Imaging Spectroradiometer
ML	Machine Learning
MSG	Meteosat Second Generation
NASA	National Aeronautics and Space Administration
NRX	Nystrom RX
NOAA	National Oceanographic and Atmospheric Administration
PCA	Principal Component Analysis
PR	Precision Recall
PROBA	Project for On-Board Autonomy
RBIG	Rotation-Based Iterative Gaussianization
RBF	Radial Basis Function
RKHS	Reproducing Kernel Hilbert Space
RFF	Random Fourier Features
RS	Remote Sensing
ROC	Receiver Operating Characteristic
ROI	Region of Interest
RX	Reed-Xiaoli (detector)
SAM	Spectral Angle Mapper
SEVIRI	Spinning Enhanced Visible and Infrared Imager
SPOT	Satellite Pour l'Observation de la Terre
SVD	Singular Value Decomposition
SVDD	Support Vector Domain Description
ORF	Kernel Orthogonal Random Features
TD	Target Detection
VHR	Very High Resolution
WTC	World Trace Center

ABSTRACT

This Thesis deals with the relevant problems of detecting changes, anomalous changes and anomalies in remote sensing images for Earth observation (EO). The Thesis' main objective is to develop, characterize, improve, implement and apply novel and robust detectors under two statistical approaches to estimate patterns from data; kernel methods and multivariate Gaussianization.

On the one hand, kernel machines constitute a proper framework to develop detection algorithms, to accommodate multi-source data, model complex distributions, to cope with high-dimensional data, and can be engineered to the particular EO signal characteristics, such as unevenly sampled time series and missing data, non-Gaussianity, and non-stationary processes. Current anomaly detection algorithms are typically challenged by either accuracy or efficiency. More accurate nonlinear detectors are typically slow and do not scale well to millions of pixels. Kernel methods provide a consistent and well-founded theoretical framework for developing nonlinear techniques and have useful properties when dealing with low-to-moderate amount of (potentially high dimensional) training samples. One of the specific objectives of this Thesis is to improve kernel detectors, both by developing automatic procedures and fast kernel models. The proposed methods achieve relevant results in terms of detection accuracy, reducing the false alarm rates, and minimizing the computational cost. However, the kernel methods present a problem with parameter settings. A new framework to deal with kernel parameter adjustment is also implemented. This setting is based in explicit density estimation. The rotation-based iterative Gaussianization (RBIG) approach is a non-parametric method that can be used for density and information theoretic measure estimation, long-standing problems in statistics and machine learning. The RBIG method is unsupervised and it is proposed for for detecting anomalies and changes in remote sensing images. The methodology transforms arbitrarily complex multivariate data into a multivariate Gaussian distribution. Therefore, one can estimate the probability at any point of the original domain and take pixels with low estimated probability as anomalies.

In this Thesis, several very challenging problems are addressed. The anomalous changes detection at very high spatial resolutions implicitly includes the change detection

and the anomaly detection settings as particular cases. The performance of all algorithms was studied in a representative number of multispectral and very high resolution satellite images such as AVIRIS, Sentinel-2, WorldView-2, MODIS, Quickbird and Landsat8, as well as in a wide range of situations involving droughts, wildfires, floods, and urbanization. The methods are based mainly in estimating either distances or probabilities. The distance setting is represented by the well-known Reed-Xiaoli (RX) detector as well as the Cook's distance detector, which are extended to cope with non-linearities for anomaly change detection. The RX method is improved by proposing kernelization of the elliptically-contoured distribution. On the other hand, the Cook's distance is extended by a novel kernelized version to address anomalous change problems whereas the random Fourier features and Nyström implementations help us to approximate the kernel solution and improve the computationally efficiency. The probabilistic setting is represented by means of the RBIG methodology, which is able to describe any multivariate distribution, making an efficient use of memory and computational resources, while being a parameter-free method for density estimation. It has been anticipated that detection of rare, unexpected changes and events under the developed statistical framework will constitute the stepping stone for the monitoring and protection of areas that are difficult to access. Our proposed techniques demonstrated good performance over the state-of-the-art approaches and will contribute to address the challenges around anomaly and change detection in remote sensing in the future.

RESUMEN

Esta Tesis va enfocada principalmente a tratar los problemas de detección de anomalías y cambios en el ámbito de Observación de la Tierra usando imágenes satelitales. El objetivo principal es implementar detectores para hacer frente a las diferentes adversidades que están presentes en la naturaleza. Básicamente nos referimos a aquellos eventos o situaciones que se consideran atípicos o fuera de lo normal como es el caso de las sequías, inundaciones, incendios forestales, urbanizaciones y otros ejemplos que a menudo suelen aparecer durante la monitorización de la Tierra. En la actualidad la mayoría de los algoritmos que tratan la detección de anomalías y cambios anómalos suelen ser cuestionados por la precisión o la eficiencia a la hora de detectar estos eventos. Para hacer frente a estos problemas, esta Tesis se basa en dos marcos principales para desarrollar, mejorar e implementar detectores robustos. El eje central de la Tesis se basa en los métodos Kernel que proporcionan un marco teórico consistente y bien fundamentado para el desarrollo de técnicas no lineales y presentan propiedades útiles cuando se trata de un número bajo de muestras de entrenamiento en datos de alta dimensionalidad. Uno de los problemas a los que nos enfrentamos con estos métodos es el alto coste computacional debido a la gran cantidad de datos que presentan las imágenes satelitales. De aquí se deriva otro de los objetivos de esta Tesis: desarrollar modelos automáticos, rápidos y eficientes basados en aproximaciones del Kernel y que además superen tanto en predicción como en precisión a los métodos lineales. Por otra parte, se ha utilizado también otro marco teórico basado en la estimación explícita de la densidad. Esta parte se enfoca en la necesidad de desarrollar algoritmos de detección entrenados de manera no supervisada, ya que los métodos basados en Kernel y sus aproximaciones necesitan ajustar de forma manual o mediante la validación cruzada sus parámetros principales. Se utilizará la Gaussianización iterativa basada en la rotación, el cual es un modelo no paramétrico. Este método se utiliza de manera no supervisada para la detección de cambios y anomalías en las imágenes de teledetección. La técnica de Gaussianización multivariante permite estimar con precisión las densidades multivariantes, un problema clásico en estadística y el aprendizaje automático sobre todo cuando los datos tienen una gran dimensionalidad. En síntesis, esta metodología transforma datos multivariados arbitrariamente complejos en una distribución gaussiana multivariada.

En esta Tesis, abordaremos un problema muy ambicioso: la detección de cambios anómalos a muy altas resoluciones espaciales que implícitamente incluye la detección de cambios y la de detección de anomalías como casos particulares. Se estudió el rendimiento de todos los algoritmos en un número representativo de imágenes satelitales multiespectrales y de muy alta resolución como AVIRIS, Sentinel-2, WorldView-2, MODIS, Quickbird y Landsat-8, así como en una amplia gama de situaciones relacionadas con sequías, incendios forestales, inundaciones y urbanización. Estos métodos se basan principalmente en la estimación de distancias y probabilidades. Los modelos basados en distancia están representados por el conocido Reed-Xiaoli (RX) y su familia de detectores, así como por la distancia de Cook y sus aproximaciones. Ambos enfoques hacen referencia a versiones lineales y no lineales para la detección de anomalías y cambios anómalos en imágenes de teledetección. La familia de los métodos RX es extendida a su versión no lineal mediante el uso de kernels de forma que es capaz de mejorar la precisión de la detección con respecto a los métodos lineales originales. Por otra parte, la distancia de Cook se extiende mediante el uso de kernels para abordar los problemas de cambios anómalos. Además, se utilizan aproximaciones del Kernel basadas en el método de características aleatorias de Fourier y el método de Nyström que nos ayudan a mejorar la eficiencia, el coste computacional y la precisión del modelo. En el caso del método basado en estimación de probabilidades se utilizó la metodología de Gaussianización ya que permite estimar con precisión las densidades multivariantes. El método se fundamenta en la idea de la Gaussianización multivariada, que consiste en buscar una transformación que convierta un conjunto de datos multivariados a un dominio en el que los datos mapeados sigan una distribución normal multivariada. Por lo tanto, aplicando la fórmula del cambio de distribucion bajo transformaciones, el modelo permite estimar la probabilidad en cualquier punto del dominio original haciendo un uso eficiente de los recursos de memoria y de computación. Se implementó esta estimación para determinar que los píxeles de baja probabilidad estimada se consideran anomalías. Además, este método no necesita ajustar ningún parámetro lo que lo convierte en una modelo no supervisado. Se demuestra la eficiencia del método en experimentos que implican tanto la detección de anomalías como la detección de cambios en diferentes conjuntos de imágenes de satélites. Para la detección de anomalías proponemos dos enfoques. El primero utilizando directamente la Gaussianización iterativa basada en rotación (RBIG) y el segundo utilizando un modelo híbrido que combina la Gaussianización y el método Reed-Xiaoli (RX) que habitualmente es utilizado en la detección de anomalías.

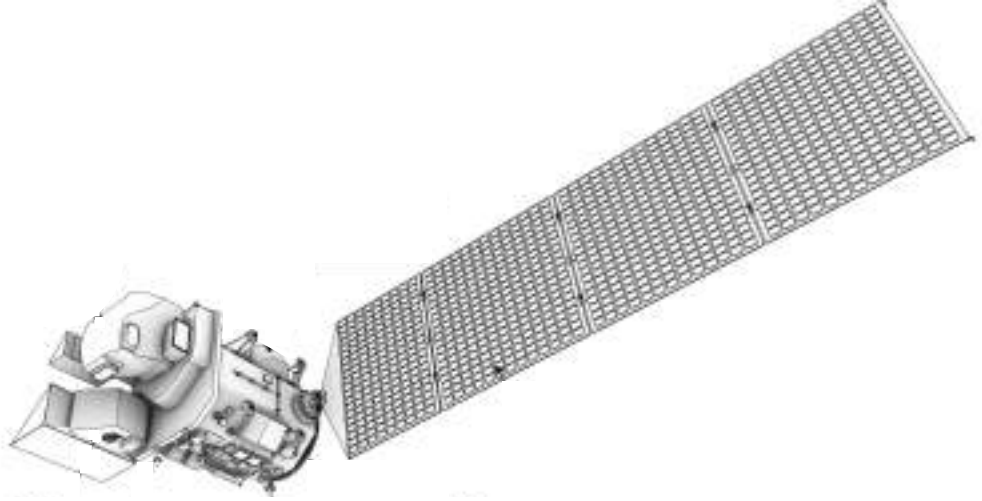
En resumen, esta Tesis se ha enfocado en la detección de cambios naturales, inesperados

y poco frecuentes (anómalos) entre pares de imágenes tomadas por sensores satelitales, así como la detección de anomalías en una sola imagen (satelital) bajo técnicas estadísticas. Se implementan nuevos y avanzados métodos de aprendizaje automático basados en métodos Kernel, desarrollamos aproximaciones eficientes y se utilizan métodos de estimación de densidad tanto supervisados como no supervisados. Todos estos métodos constituirán un paso importante para la monitorización y protección de las zonas de difícil acceso sin la necesidad de la presencia del ser humano, reduciendo así en gran proporción el coste económico que este pudiese causar.

RESUM

Aquesta tesi tracta dels problemes de detecció de canvis, canvis anòmals i anomalies en imatges de teledetecció per a l'observació de la Terra. L'objectiu principal de la tesi és desenvolupar, caracteritzar, millorar, implementar i aplicar detectors estadístics seguint dos marcs fonamentals: mètodes nucli i modelització probabilística. D'una banda, les màquines nucli constitueixen un marc adequat per desenvolupar algoritmes de detecció, perquè poden tractar dades de múltiples fonts, modelar distribucions complexes, processar dades d'alta dimensió i poden adaptar-se a les característiques particulars dels senyals d'observació de la Terra, com ara un mostreig no-uniforme, no gaussianitat i processos no estacionaris. Els algoritmes de detecció d'anomalies actuals solen ser qüestionats per la seua manca de precisió o eficiència davant estos problemes. Mentre que els detectors no lineals solen ser més precisos, també son més lents i no escalen bé a problemes amb milions de píxels. Els mètodes nucli proporcionen un marc teòric consistent i ben fonamentat per al desenvolupament de tècniques no lineals i tenen propietats bastant útils quan es tracta d'un nombre baix de mostres d'entrenament (potencialment d'alta dimensió). Un dels objectius específics d'aquest treball és desenvolupar models automàtics i ràpids amb un bon rendiment en termes de precisió en la detecció i eficàcia computacional. L'objectiu és obtenir resultats rellevants, reduint les taxes d'alarma falsa i minimitzant el cost computacional dels algorismes no lineals. Tanmateix els mètodes nucli presenten un problema amb la configuració dels paràmetres. Un marc estadístic alternatiu considera la estimació explícita de la densitat i prendre el concepte d'anomalia o de canvi com aquell més sorprenent, el de menor probabilitat. En esta tesi presentem un algoritme basat en Gaussianització anomenat RBIG: és un mètode no paramètric que es pot utilitzar per a estimar densitats multi-variades i també per tal d'estimar mesures de la teoria de la informació, ambdós problemes encara no resolts als camps de la estadística i l'aprenentatge automàtic. El mètode RBIG és no-supervisat i el proposem per detectar anomalies i canvis en les imatges de teledetecció. La metodologia transforma dades multi-variants arbitràriament complexes en una distribució gaussiana multi-variant. A més, es pot estimar la probabilitat en qualsevol punt del domini original, cosa que significa que els píxels amb baixa probabilitat estimada es consideren anomalies.

En aquesta tesi abordarem diversos problemes molt difícils: la detecció de canvis anòmals a resolucions espacials molt altes, i què inclouen implícitament la detecció de canvis i la detecció d'anomalies com a casos particulars. Estudiem el rendiment de tots els algoritmes en un nombre representatiu d'imatges de satèl·lit multi-espectrals i hiperespectrals, de molt alta resolució espacial com AVIRIS, Sentinel-2, WorldView-2, MODIS, Quickbird i Landsat8, així com en una àmplia gamma de situacions relacionades amb sequeres, incendis forestals, inundacions i urbanització. Els mètodes es basen principalment en l'estimació de distàncies i probabilitats. La configuració de la distància està representada pel conegut detector Reed-Xiaoli (RX), així com la distancia de Cook: ambdós mètodes son generalitzats al cas no lineal amb kernels i aplicats a la detecció de canvis anòmals. El mètode RX s'estén amb una família de detectors anòmals on la distribució contornada el·lípticament no lineal és capacc de millorar la precisió de detecció respecte a la lineal. D'altra banda, la distància del Cook s'estén amb una nova versió kernelitzada per abordar problemes de canvis anòmals, mentre que les característiques aleatòries de Fourier i les implementacions de Nyström ens ajuden a aproximar la solució del nucli i així millorar l'eficiència computacional. L'aproximació probabilística està representada pel RBIG, que és capacc de descriure qualsevol distribució multi-variant i aquest mètode fa un ús eficient de la memòria i els recursos computacionals, convertint-se en un mètode lliure de paràmetres. Les tècniques proposades en aquesta Tesi Doctoral han demostrat un bon rendiment respecte als enfocaments d'última generació i contribuiran a abordar els reptes relacionats amb la detecció d'anomalies i la detecció de canvis en la teledetecció futura.



1. INTRODUCTION

Contents

This Chapter highlights the importance of Earth Observation (EO) to cope with the natural and anthropogenic events (changes, anomalies) using remote sensing satellite images. The main concepts concerned in the thesis will be reviewed, as well as the main goal: developing and improving algorithms for change detection (CD), anomalous change detection (ACD) and anomaly detection (AD) in RS images for Earth Observation.

1.1 Introduction to Earth Observation

The Earth is a highly complex and evolving system. Natural events and human activities have precipitated an environmental crisis on Earth. Quantification and understanding of both the natural and anthropogenic impact on the Earth's system is matter of current and intense research, and one of the biggest challenges in nowadays science. By using remote sensing sensors it will be possible to identify materials on the land cover. The problem of global warming has also deep implications of relevant societal, environmental, and economical values, given the rapidly growing demand of biofuels and food (IPCC, 2012). Undoubtedly, there is an urgent need to provide quantitative monitoring tools of the Earth system processes. Advances in data exploitation using remote sensing satellite images and products have allowed us to improve predictions and understanding of extreme events occurring across the Earth's surface.

Earth observation (EO) data analysis is nowadays a mature field of science, where many real-life applications of societal value are developed. Perhaps one of the most important problems in EO data processing is the detection of anomalous changes in land-cover classes or spatial-temporal extreme events in satellite images and products. Actually, anomaly detection and anomalous changes such as precipitation events, heat waves, latent fires, droughts, floods and urbanization are increasingly perceived as key players in the Earth system, and are expected to increase in the wake of climate change (IPCC, 2012). These problems involve complex, heterogeneous, multi-modal, multi-source and structured data: problems arise at very high resolution (VHR) or larger (kilometric) spatial resolutions, and also at different spectral and temporal resolutions. Measurement, analysis and interpretation of the spectrum is a very vast field in spectroscopy (Danson & Plummer, 1995; Liang, 2004; Lillesand et al., 2014; Richards & Jia, 1999; Ustin, 2004) . In the field of EO, *remote sensing* is performed by sensors typically onboard either satellite or airborne platforms thus recording reflected or emitted electromagnetic energy from Earth's surface. On the other hand, EO remote sensing can be also achieved with "near-surface" approaches using RGB (red, green, blue) cameras placed on selected sites for monitoring

floods, wildfires and many kind of the anomalous events.

Remote sensing sensors can be classified either as passive or active. Passive sensors record reflected Earth's surface energy that was emitted from the sun. The solar radiation received by a sensor is detected at different wavelengths and the resulting spectral signature is used to identify a specific material. There are a variety of passive sensors such as: infrared, coupled devices, radiometers, and multispectral and hyperspectral sensors (Shaw & Manolakis, 2002). Whereas, active sensors record reflected Earth's surface electromagnetic energy which was previously emitted by themselves (Mott, 2007; Wang, 2008). In particular, it will be focused on multi and hyper spectral passive sensors. Even is valid to highlight that the methodologies are generic to any sensor or even any data set not necessarily remote sensing, e.g. time series, data science, big Data, etc. However the proposed methodologies could be applied straightforwardly to active sensors data.

The information conveyed by remotely sensed platforms is tied to the sensor characteristics and satellite capabilities (Lillesand et al., 2014). Spatial resolution gives the image pixel size (ranging from centimeters to kilometers), spectral resolution provides data at different spectral wavelengths (ranging from solar to thermal spectrums) and also information about spectral width, temporal resolution relates both the acquisition date and the frequency of acquisitions (ranging from a day to decades), and eventually spatial extent covers the ground area detected by the sensor's field of view (up to the entire Earth). Consequently, the most suitable sensor/platform must be selected depending on the needs of the application (Benz et al., 2004).

1.1.1 Remote Sensing at global scale

One of the main goals in remote sensing (RS) is to study the detection of anomalies in global scale processes to improve our knowledge about how Earth functions. This has been partly achieved employing different methods to determine and evaluate the behavior using pairs or time series images. RS has been used to monitoring of the areas with difficult access. Different kind of satellites are used, such as with medium resolution (i.e. kilometric spatial resolution) sensors like: Moderate Resolution Imaging Spectroradiometer (MODIS, 1999 up to date) (Justice et al., 1998), the Système Pour l'Observation de la Terre (SPOT-V, 1999 to May 2014) (Pasquier & Verheyden, 1998) and the Project for On-Board Autonomy (PROBA-V, June 2014 up to date) (Sterckx et al., 2014). Similarly, the European Organization for the Exploitation of Meteorological Satellites (EUMETSAT) provides information for vegetation monitoring from the Spinning Enhanced Visible and Infrared Imager (SEVIRI) onboard Meteosat Second Generation (MSG). Also, there are

many sensor with very high resolution to detect with better precision the anomalies. The National Oceanographic and Atmospheric Administration/ Advanced Very High Resolution Radiometer (NOAA/AVHRR) sensor is used to monitor vegetation (Townshend, 1994). Among the most recent developments in remote sensing are high spatial resolution sensors, those of high spectral resolution and finally the use of Synthetic Aperture Radar (SAR). Could be highlight the Sentinel-2 satellites, managed by ESA (European Space Agency), and Landsat 8, managed by NASA/USGS (National Aeronautics and Space Administration / United States Geological Survey). KOMPSAT 3 and KOMPSAT 3A were launched to continue KOMPSAT 1 and KOMPSAT 2 missions accordingly dedicated to natural disasters activities. Both satellites are fitted with a pushbroom imager, AEISS (Advanced Earth Imaging Sensor System) with a spatial resolution of 0,5 m/pxl for KOMPSAT 3 and 0,4 m/pxl – for KOMPSAT 3A which was enhanced with IIS (Infrared Imaging System) to operate within the Mid-Wavelength Infrared region of 3 – 5 μm with high thermal resolution.

1.1.2 Remote Sensing at local scale

Recent advances in remote sensing platforms, sensors, statistical models and as well as the enormous amount of data available have provided new challenges and possibilities in remote sensing image processing (Campos-Taberner et al., 2016). Especially, in the context for events detection they have entailed in many studies of anomaly and change detection. Accurate and timely information at high spatial resolution of crop condition and status is critical for crop management. For example, the Landsat Data Continuity Mission (LDCM) (Roy et al., 2014) and the recently European Sentinel-2 Mission (Drusch et al., 2012.) provide free EO high-resolution (HR) information for a wide variety of land applications (Malenovský et al., 2012) including crop monitoring. Several HR data are also available from different initiatives and platforms such as Worldview or RapidEYE. However, they are not free of charge which limits its usability in continuous and long term applications.

1.1.3 The detection problem in remote sensing

Many remote sensing applications do not need the definition of many classes of interest. Actually, very often, they only need to discriminate a single class from the rest, i.e. the background (Ahlberg & Renhorn., 2004). Sometimes, there is no knowledge about the spectral signature of the class of interest, but it is known that this signature is different from the rest. In this case, a model is built by looking for signatures that deviate from a model of

the background. Large deviations from the mean (background) are called leverage, and the models based on this principle are referred to as anomalous change detectors. Anomalous change detectors can be very difficult to calibrate, especially when there is a strong spectral variability in the image. Detecting anomalous changes in different pairs of images or time series is of paramount relevance for Earth monitoring. The field of change detection is vast and many approaches are available in the literature (Lu et al., 2004; Mas, 1999). Change detection boils down to identify the abnormality buried in the background. This Thesis is focused on the anomalous change detection (ACD) problem too (Theiler et al., 2010). Anomalous samples are essentially deviations from what is typical or expected, but a formal specification of anomalousness is complex and elusive. Unlike pervasive changes, anomalous changes and events are rare and concentrate in the tails of the distributions, which are not easy to analyse and characterize. The ACD problem is ill-defined, e.g. quoting James Theiler: *"If individual anomalies resist definition, how can we expect to write down a probability distribution for all anomalies?"* (Theiler, 2014). In fact, under probability density function (PDF) transformations, one should be able to focus only on the Jacobian to assess asymmetries, and hence characterize the anomalousness of the change. This is technically difficult as the transformation is seldom accessible. Consequently, a mathematical framework that avoids explicit distribution modelling is needed.

Current EO statistical data analysis faces an important problem. The world is witnessing an ever increasing amount of data gathered with current and upcoming EO satellite missions, such as the ESA Sentinels and the NASA A-Train satellite constellations. With the super-spectral Copernicus' Sentinel-2 (Drusch et al., 2012.) and Sentinel-3 (Donlon et al., 2012.) missions, as well as the planned EnMAP (Stuffler et al., 2007), HyspIRI (Roberts et al., 2012) PRISMA (Labate et al., 2009.) and ESA's FLEX (Kraft et al., 2012.) imaging spectrometer missions, an unprecedented data stream for land monitoring becomes available. At the same time, very high resolution (VHR) sensors allow us to monitor the planet at local and regional scales, but with an unprecedented high spatial detail: Wordlview-2 and the recent Worldview-3 (Longbotham et al., 2014), for instance, permit detecting very fine (sub-meter) spatial details in the images, and allow for realistic applications, such as the popularized Google Maps. Currently, the new sensors increase the resolution either spatial or spectral, as well as the revisit time. This leads to a big data processing problem, where algorithms should not only be accurate but scalable to the big data deluge. The company Planet has three satellite constellations (SkySat, Dove, and RapidEye) with more than 150 satellites supplying imagery and derived products over the entire Earth at medium and high resolution with high repeat frequencies. Besides, with

unrivaled accuracy, agility and collection capability, MAXAR offers customers around the world affordable access to the highest quality vision of their world through its high-resolution sensors. It should be highlighted that the use of drones have been popularized too in military, industrial, civil or security fields. Some of the main companies are ADTS Group and Aerial Insights from Spain. All this data influx requires advanced anomaly detection techniques, which should be accurate, robust, reliable and fast.

1.2 Detection of anomalies, changes and anomalous changes

Here it is aimed to formally distinguish the main concepts of anomaly detection, change detection and the detection of anomalous changes in the context of remote sensing. The main characteristics, the literature and the methods used in each field are point out and reviewed.

1.2.1 Change detection

The field of **change detection** (CD) methods is extensive and many approaches are available in the literature (Lu et al., 2004; Manolakis et al., 2003; Manolakis & Shaw, 2002; Radke et al., 2005). This approach can be organized in three types of products (Coppin et al., 2004; Singh, 1989): 1) binary maps, 2) different types of change detection, and 3) full multiclass change maps. Each of them can be obtained using different sources of information extracted from the initial images at time instants t_1 and t_2. Unsupervised CD has been widely studied due to the high relationship with most applications: i) the speed for obtain the change map and ii) the lack of labelled information in the applications. The problem of change detection deals with identifying transitions between a pair (or a series) of co-registered images (Radke et al., 2005; Singh, 1989). Change detection in remote sensing images is of paramount relevance because it automatizes traditional manual tasks in disaster management (floods, droughts and wildfires) and it helps in designing development and settlement plans as well as in urban and crop monitoring. Multitemporal classification and change detection are very active fields nowadays because of the increasingly available complete time series of images and the interest in monitoring changes occurring on the Earth's cover due to either natural or anthropogenic activities. Complete constellations of civil and military satellites sensors currently provide high spatial resolution and high revisiting frequency. The Copernicus' Sentinels[1] or NASA's A-train[2] programs are producing near real-time coverage of the globe. NASA is currently producing a Harmonized Landsat

[1] http://www.esa.int/esaLP/SEMZHM0DU8E_LPgmes_0.html
[2] http://www.nasa.gov/mission_pages/a-train/a-train.html

Sentinel-2 (HLS) data set, which can be used for monitoring agricultural resources with an unprecedented combination of 30 m spatial resolution and 2-3 days revisit. In parallel, new commercial satellite missions are being deployed to provide multispectral data at both high spatial and high temporal resolution. For example, the PlanetScope constellation by Planet Labs, Inc. can provide 5 m data daily for sites requested by the client, and the recently announced UrtheDaily constellation, specifically designed for operational agricultural applications, will acquire S2-like data also at 5-m spatial resolution and with full global coverage every day. It goes without saying that closed-range applications using drones and all kind of unmanned automated vehicles (UAVs) also challenge the field of automatic change detection. All in all, automatic image analysis in general and change detection in particular are becoming necessary in the current era of data deluge.

However, the lack of labeled information makes the problem of detection more difficult and thus unsupervised methods typically consider binary change detection problems only. In the last decade, change vector analysis (CVA) techniques have been widely applied: CVA converts the difference image to polar coordinates and operate in such representation space (Bovolo & Bruzzone, 2007; Malila, 1980). In (Dalla Mura et al., 2008), morphological operators were successfully applied to increase the discriminative power of the CVA method. In (Bovolo, 2009), a contextual parcel-based multiscale approach to unsupervised CD was presented. Traditional CVA relies on the experience of the researcher for the threshold definition, and is still on-going research (Chen et al., 2010; Im et al., 2008). The method has been also studied in terms of sensitivity to differences in registration and other radiometric factors (Bovolo et al., 2009). Another interesting approach based on spectral transforms is the multivariate alteration detection (MAD) (Nielsen, 2006; Nielsen et al., 1998), where canonical correlation is computed for the points at each time instant and then subtracted. The method consequently reveals changes invariant to linear transformations between the time instants. Radiometric normalization issues for MAD has been recently considered in (Canty & Nielsen, 2008), and nonlinear extensions have been also realized via kernel machines (KM) (Gómez-Chova et al., 2011; Nielsen, 2011). Other approaches based on KM have proposed to use dimensionality reduction via principal components (Ding et al., 2010) or slow features (Wu et al., 2017) of the difference image. Clustering has been used in recent binary CD. In (Celik, 2009), local PCAs are used in sub-blocks of the image, followed by a binary k-means clustering to detect changed/unchanged areas locally. Also, kernel-based clustering has been investigated in (Volpi et al., 2012, 2010), where kernel k-means with optimized parameters use an unsupervised cost function to separate two groups. Finally, unsupervised neural networks

(a) (b) (c) (d)

Figure 1.1: The image showcase an aerial view of the rice-field in the Albufera Natural Park (Valencia). a) image corresponding to rice crop at planting time, b) image corresponding to rice crop at harvest time, c) image contains a anomalous change (black square) and d) image correspond to the anomalous location, the interest region is highlighted with a green square around it.

have been introduced for the binary CD (Pacifici et al., 2010, 2009). In (Ghosh et al., 2007), a Hopfield neural network, where each neuron is connected to a single pixel and used to improve neighborhood relationships. Recently, it has been studied to include deep convolutional neural networks (Liu et al., 2018; Ouyang et al., 2017; Zhang & Zhang, 2014).

1.2.2 Anomalous Change Detection

A related field of investigation in this direction is the so-called **anomalous change detection** (ACD) (Theiler, 2008): in this field, one looks for changes that are interestingly anomalous in multitemporal series of images, and tries to highlight them in contrast to acquisition condition changes, registration noise, or seasonal variation. Figure 1.1 illustrates the difference between CD and ACD scenarios using remote sensing images. Changes between two images can be differentiated in regular and anomalous. Regular changes are usually defined by cyclical time patterns, for instance, the change in the vegetation's greenness with the passage of the year's season, exemplified between (a) and (b) images. On the contrary, an anomalous change is any alteration of the scene that is outside of what is normally expected: for example, the emergence of the black square between images (a) and (c). While applying CD and ACD algorithms on images (b) and (c) would get similar results. This could not be the case when applied on images (a) and (c). On the one hand, the CD algorithm would detect as a change almost all the regions in the image. This would be a good result if one is interested on detecting vegetation changes. However, one could be interested in ignoring the regular changes and detecting only the black square. In such case, an ACD algorithm would be better fitted since it ignores the brownish to

greenish changes and aims to detect as anomaly only the black square. The ACD settings are perfectly suited to the statistical theoretical framework. The concept of *anomalousness* is difficult to define concretely. Nevertheless, identifying influential points in multivariate data distributions is an active field of research in statistics, information theory, and machine learning. Main applications involve characterization of distributions, detection of anomalies, extremes and changes, and robustness assessment, just to name a few (Theiler, 2013; Theiler & Wohlberg, 2012). Detection of such influential points has also relevant applied implications for climate, health and social sciences, and in a wide diversity of engineering, communications and computer science problems. Since the seminal work of Cook (Cook, 1977), many diagnostic measures have been introduced, most of them relying on the adoption of parametric or non-parametric models, from which residuals and then leverages are estimated. Diagnostics have been introduced for both *parametric models*, such as linear regression (Cook & Weisberg, 1980; Snee, 1983), penalized (ridge) regression (Hoerl & Kennard, 1970), sparse regression models like LASSO (Tibshirani, 1996), and *non-parametric models* such as spline smoothing (Choongrak & Barry, 1996; Eubank, 1985; Silverman, 1986) and polynomial regression (Kim et al., 2001). Extensions have considered *semi-parametric models* (Fung et al., 2002), longitudinal regression (Bae et al., 2008), as well as generalized linear and Cox proportional hazard models (Zhu et al., 2015, 2009, 2001).

An adequate model assumption and specification is crucial and has many theoretical and applied implications. The main problem is to select a flexible model that can capture nonlinear relations, but also provides high detection power and computational efficiency. All these are relevant aspects to consider for the diagnostic measure, for which many methods have been proposed. In recent years, kernel methods have been widely adopted as an appropriate framework for model development in machine learning for classification, regression, hypothesis testing and dimensionality reduction (Rojo-Álvarez et al., 2017; Schölkopf & Smola, 2002). Kernel methods allow to derive flexible nonlinear and non-parametric models, are intrinsically regularized and are endorsed with solid mathematical properties. This has allowed to define diagnostics based on leveraging for the kernel ridge regression (KRR) method (Shawe-Taylor & Cristianini., 2004). However, despite the excellent modeling performance of KRR, the direct definition of leverage scores based on KRR implies a huge computational cost and the lack of a practical out-of-sample estimate (Alaoui & Mahoney, 2015; McCurdy, 2018; Rudi et al., 2018). This hampers its adoption and usefulness in real practice. The interest to find anomalous changes in scenes is very broad, and many methods have been proposed in the literature,

ranging from equalization-based approaches that rely on whitening principles (Mayer et al., 2003), to multivariate methods that extract distinct features out of the change (difference) image (Arenas-Garcia et al., 2013) and that reinforce directions in feature spaces associated with noisy or rare events (Green et al., 1998; Nielsen et al., 1998), as well as regression-based approaches like in the chronochrome (Schaum & Stocker, 1997), where a regression model approximates the next incoming image and big residuals are associated with anomalies.

This Thesis starts with the Cook's distance approach for ACD (second chapter). This method was develop by Dennis Cook in 1977. Our main goal for this setting is recovering this traditional vision and developing a nonlinear kernel-based extension by using kernel regression in a reproducing kernel Hilbert space (RKHS). Noting the high computational cost that a naive formulation has, the random Fourier features (Rahimi & Recht, 2007) and the Nyström method (Williams & Seeger, 2001; Zhao et al., 2017) were introduced for improving the efficiency. Both approaches allow us to compute residuals (Touati et al., 2020) and leverages of the KRR *explicitly* in RKHS while Nyström method also provides implicit regularization capabilities. Essentially, Nyström method approximates the large kernel matrix by a much smaller low-rank matrix. Although, best low-rank approximation is obtained by Singular value decomposition (SVD), it is computationally expensive. Per contra, Nyström method achieves low-rank approximation with considerably higher computationally efficiency (Kumar & Schneider, 2017; Yang et al., 2012). Thus, low-rank methods are already used in various geoscience applications (Cao et al., 2019; Sun et al., 2020). In this context, empirically was validated the accuracy-speed trade-off of our methods in challenging problems of ACD using big satellite image datasets. A fast kernelized Cook's distance estimates to evaluate change were proposed on the chronochrome approach (Longbotham & Camps-Valls, 2014; Theiler et al., 2010). Essentially, the Cook's distance is defined as the sum of all changes in the regression model when a particular observation is removed. To our knowledge, the Cook's distance has not been used in image ACD in a chronochrome setting, most probably because the linearity assumption is too rigid in general CD problems where nonlinear processes, such as illumination changes, occlusion, backscattering and sensor distortions, occur in the wild. It is important to remark that in this anomalous change detection approach the aim is detecting important (extreme) changes, i.e. not pervasive changes related to, for example, illumination conditions. Therefore, this refers to anomalies among two images, leverage points or changes interchangeably.

Additionally, the Cook's distance for out-of-sample points requires the evaluation of

as many models as test samples, which makes the technique very costly and definitely impractical. A remedy is provided to both problems by (1) defining Cook's distance for the nonlinear KRR model and by (2) introducing two computationally efficient kernel approximations. The proposed methods are simple, computationally very efficient in both memory and processing costs, and achieves improved detection compared to standard approaches. The results are showed in a set of real anomalous change detection problems with pairs of multispectral satellite images acquired by different sensors (Quickbird, Sentinel-2) and involving different changes of interest (floods, wildfires, urbanization and droughts). The datasets are large scale and thus a good testbed for our approach.

On the other hand, the study based on the detection of anomalous changes (in chapter 3) is continued following the work in (Theiler & Perkins, 2006). This formalized the field by introducing a framework for ACD, which assumes Gaussianity, yet the derived detector delineates hyperbolic decision functions. Even though the Gaussian assumption reports some advantages (e.g. tractability and generally good performance) it is still an *ad hoc* assumption that it is not necessarily fulfilled in practice. This is the motivation in (Theiler et al., 2010), where the authors introduced elliptically-contoured (EC) distributions that generalize the Gaussian distribution and proved more appropriate to modeling fatter tail distributions and thus detect anomalies more effectively. The EC decision functions are point-wise nonlinear and still rely on second-order feature relations. Recent advances in ACD have considered methods robust to pixel misregistration (Theiler & Wohlberg, 2012) and sequences of several images (Theiler & Adler-Golden, 2008). The theory of reproducing kernels is used to implement the nonlinear version of the Gaussian assumption and the elliptically contoured distribution. Making an extension of the family of ACD methods presented in (Theiler et al., 2010).

1.2.3 Anomaly detection

The use of hyperspectral imagery to perform target detection and recognition has been widely investigated and has proven valuable in many applications including search-and-rescue operations, border surveillance, and mine detection. This kind of approach can be split into two main applications: classification and target detection (TD). TD can be considered as a binary classification problem: the aim is to classify the image into the target class and the background class. However, regardless of the application, the general goal of TD is to detect small rare objects that constitute a very small fraction of the area of the background in which they are embedded.

In this Thesis, one of the main topics is anomaly detection (chapter 4 and 5), which can

be viewed as a particular case of TD in which no a priori information about the spectrum of the targets of interest is provided. Anomaly detection (AD), as a remote sensing (RS) research topic, is gaining importance because of the need for processing large number of images that are acquired from satellite and airborne platforms (Stein et al., 2002). The goal of anomaly detection is to find objects in the image that are anomalous with respect to the background. There is not a particular way to define an anomaly, which can be generally identified as an observation that deviates in some way from the rest of the distribution. Even the background itself can be identified in different ways, such as a local neighborhood surrounding the observed pixel or as a larger portion of the image. In addition, it is important to mention different kind of background characteristic or scenarios such as: crop stress location in agriculture applications, infected trees in forestry, rare minerals in geology applications, or man-made objects and vehicles in defense and surveillance applications.

Anomaly detectors (ADs) assume no a priori knowledge about the target spectral signature and simply explore the data cube to find those pixels whose spectrum significantly differs from the background. As previously stated, the anomalies of interest can change, depending on the particular application. Furthermore, a complex scenario may contain a number of anomalies which do not necessarily represent a target of interest, and also, small regions of background pixels can be detected as anomalies. As a matter of fact, since ADs do not use any a priori knowledge, they cannot distinguish between legitimate anomalies and detections that are not of interest. Therefore, the detected anomalies may include man-made targets, natural objects, image artifacts, and other interferers. In fact, anomaly detection is often a first step in the analysis of the scene, providing the regions of interest (ROIs) that may contain potential targets; ROIs can then be explored with spectral matching algorithms or spectroscopy techniques or can be used to cue higher spatial resolution sensors for target classification and matching.

Among the many dectector algorithms found in the literature, the Reed-Xiaoli (RX) detector (Reed & Yu, 1990a) is widely used due to its good performance and simplicity. The RX detector determines target pixels that are spectrally different from the image background based on the Mahalanobis metric. For RX to be effective, anomalous targets must be sufficiently small compared to background and is assumed to follow a Gaussian distribution. However, it has been shown that the Gaussian distribution assumption fails, for example, in hyperspectral images or with complex feature relations, especially at the tails of the distribution (Matteoli et al., 2010). As a result, nonlinear versions of RX have been introduced to mitigate this problem, and the kernel RX (KRX) detector was

proposed in (Kwon & Nasrabadi, 2005) to cope with complex and nonlinear backgrounds. However, the KRX algorithm has not been widely adopted in practice because, being a kernel method, the memory and computational cost increase, at least quadratically, with the number of pixels. This poses the perennial problem of accuracy versus usability in nonlinear detectors in general and kernel anomaly detectors in particular. In the literature, *local* and *global* RX-based detectors have been proposed.

One objective is focused on improving the space (memory) and time efficiency of the KRX anomaly detector. Kernel-based anomaly detectors provide excellent detection performance since they are able to characterize non-linear backgrounds (Camps-Valls & Bruzzone, 2009.). In order to undertake this challenge, the use of efficient techniques based on random Fourier features and low-rank approximations to obtain improved performance of the KRX algorithm were proposed. Initial efforts were focused using the random Fourier features approach in (Nar et al., 2018). In local AD (Reed & Yu, 1990a), pixels in a sliding window are used as input data. Despite their adaptation to local relations, the detection power has been shown to be low recently (Guo et al., 2016; Matteoli et al., 2010). Conversely, in global AD all image pixels are used to estimate statistics. Thereby, targets with various sizes and shapes can be detected while detection of small targets can be difficult. Finally, all the methods are used in a global setting for the sake of simplicity.

1.3 Advancing Machine Learning Detectors

Over the last few decades, a wide diversity of algorithms for anomaly, change and anomalous change detection have been introduced in remote sensing data analysis, but only a few of them made it into operational processing chains, and many of them are only in its infancy. There is an additional, more scientifically challenging, problem: the lack of adoption of a unified mathematical framework for anomaly detection. Both issues call for advances in EO data processing methods. In this context, statistical inference also known as machine learning play a fundamental role nowadays.

Anomaly and Change detection approaches

Anomalies are considered as data samples that are different to the rest of the data distribution. There are different ways to define 'different' or 'anomaly' but in the context and literature of *anomaly detection* it is typically agreed that anomalies are observations that deviate in some way from the background clutter. Several anomaly detection techniques have been presented in the literature. The techniques based on the Nearest Neighbor algorithm describe how data with similar behavior is found in the most dense neighborhoods,

while atypical samples should be far away from their nearest neighbors. Other methods, such as K-means, where data points in a distribution with large distance values from the center of their clusters are considered anomalies, are widely used in the clustering scenario. Nowadays, neural networks can do a good job in detection too. They are trained to learn the normal class and detect the anomaly class with respect to the background. One of the great applications of neural networks are based on generative models like GANs and VAEs, which just started to be applied in these problems by the remote sensing community.

On the other hand, the change detection goal is to compare two images and define if there is a variation between them. There are many approaches for change detection. The image difference methods are simple and direct, since the difference value of both images will be close to 0 and it will be considered a no change value. However, these techniques highly depend on the co-registration procedure. Techniques such as image regression will be presented throughout of this Thesis. These are based on a regression model that establishes relationships between bi-temporal images. The model can predict pixel values by using a regression function and classify as anomalous the one with higher residuals. In addition, the classification category includes post-classification comparison, spectral-temporal combined analysis, expectation–maximization algorithm (EM) change detection, unsupervised change detection, hybrid change detection, and artificial neuronal network (ANN).

Despite the vast and interesting amount of theory behind anomaly and change detection, the Thesis will be focused on two main learning paradigms that enable event detection: (1) a purely discriminative approach under the learning framework of kernel methods, and (2) a fully non-parametric approach for the harder problem of density estimation.

1.3.1 Kernel methods for anomaly and change detection

These approaches will be based on the framework of kernel learning (Shawe-Taylor & Cristianini., 2004), which has emerged as the most appropriate setting for remote sensing data analysis in the last decade (Camps-Valls & Bruzzone, 2009.). Kernel methods allow us to generalize algorithms expressed in terms of dot products to account for higher-order (non-linear) feature relations, yet still relying on linear algebra (Schölkopf et al., 1999; Shawe-Taylor & Cristianini., 2004). Kernel machines excel in low-to-moderate sized datasets, can accommodate multi-source data, model complex distributions with flexible kernel functions, cope with high-dimensional data, and can be engineered to the particular EO signal characteristics, such as unevenly sampled time series and missing data, non-Gaussianity, non-stationary processes, and non-i.i.d. (spatial and temporal) relations. Kernel methods

have been traditionally designed for classification and regression problems. However, the family of kernel methods currently expands to multi-temporal change detection (Muñoz-Marí et al., 2010), non-linear dependence estimation (Camps-Valls et al., 2010), hypothesis testing (Gretton et al., 2012), and anomaly detection (Longbotham & Camps-Valls, 2014), which constitute the playground of this thesis.

In this Thesis, kernel methods are introduced for a nonlinear extension of a full family of anomalous change detectors. In particular, it is focused on algorithms that utilize Gaussian and elliptically-contoured distribution modeling and extend them to their nonlinear counterparts based on the theory of reproducing kernels Hilbert space. In this particular case, it is illustrated the performance of the kernel family methods introduced for anomalous change detection problems. On the other hand, the kernel theory is used for developing a family of anomaly detectors. The family is based on approximations of the kernel matrices involved to improve the computational cost and the efficiency of all the detectors. The Random Fourier Feature (RFF) technique (Rahimi & Recht, 2007) defines a known mapping which takes the data held in the input space and transfers them to a new finite dimensional Euclidean space, where the problem is linearly separable and the inner product of the mapped data approximates the kernel function. Once the "linear" task is solved, a vector of fixed size is obtained, which defines a hyperplane in the new space, separating the data. Thus, the algorithm provided is computationally more efficient and, as will be shown throughout this Thesis, converges at similar speeds and to similar error scale. Orthogonal Random Features (ORF) impose orthogonality on the matrix on the linear transformation matrix $\mathbf{G} \in \mathbb{R}^{d \times d}$ which is a random Gaussian matrix. Note that one cannot achieve unbiased kernel estimation by simply replacing G by an orthogonal matrix, since the norms of the rows of G follow the $\mathcal{X}$-distribution, while rows of an orthogonal matrix have the unit norm. In addition, low-rank (Fine & Scheinberg, 2001) approximations of the kernel matrix are often considered as they allow the reduction of running time complexities to $O(r^2 n)$, where r is the rank of the approximation. The practicality of such methods thus depends on the required rank r. All these approximations are implemented throughout this Thesis. On the other hand, the kernel Cook's distance is studied for anomalous change detection in a *chronochrome* scheme, where the anomalousness indicator comes from evaluating the *statistical leverage* of the residuals of regressors between time acquisitions images. In addition, this version of the kernelized Cook's distance present a high computational cost due to remote sensing images. To fix this problem, it is proposed to approximate the kernel by means of the low rank approximation. One of the most popular techniques is the *Nyström method*, which

constructs $\tilde{\mathbf{K}}$ using a subset of "landmark" data points (Williams & Seeger, 2001).

1.3.2 Density Estimation with Gaussianization transforms

The rotation-based iterative Gaussianization is a nonparametric method for density estimation of multivariate distributions (Laparra et al., 2011). In this Thesis, it will be used as an unsupervised method for detecting anomalies and changes in remote sensing images by means of a multivariate Gaussianization methodology that allows to estimate multivariate densities accurately, a long-standing problem in statistics and machine learning. RBIG is based on the idea of Gaussianization, which consists of seeking for a transformation G_x that converts a multivariate dataset to a domain where the mapped data follows a multivariate normal distribution. Therefore, the model can estimate the probability at any point of the original domain and assume that pixels with low estimated probability are considered anomalies. An important aspect to take into account is the intrinsic characteristics of the data used to estimate the density, which has implications in the quality of the estimation. When the distribution contains even a moderate number of anomalies, an accurate density estimate will cast anomalies as regular points, i.e. non-anomalous. This vastly depends on the flexibility of the class of models used. When the model is rigid like in the RX case, this is not a problem since it cannot be adapted to the anomalies. For the KRX one can control this effect by tuning the kernel lengthscale and the regularization term, but as explained before this actually requires labeled data. This is an important aspect to take into account mostly in the anomaly detection scenario, where all data (included the anomalous samples) are used to estimate the density. Therefore, a hybrid model that combines the (too rigid) RX model with the (too flexible) RBIG model is proposed. The hybrid model first selects the data more likely not to be anomalous using RX and then uses this data to learn the Gaussianization transform with the RBIG model. This tries to avoid using anomalous data to train RBIG, which after all is intended to learn the background or pervasive data distribution. The number of data points selected as non-anomalous in the first step will define the trade-off between flexibility and rigidity. On the other hand, the same theory is applied to address change detection problems in remote sensing. It is important to note that, in this case, the data used to estimate the probability density (first image) does not contain anomalies, so the hybrid model is not needed in this approach.

1.4 Research objectives, structure and contributions

Why this Thesis?

One of the most important challenge for today's Science is to detect and attribute the causes of changes and extremes in arbitrary distributions. In the context of Earth observation data analysis, this general problem translates into the problem of automatically detecting anomalies on the land-cover at both spatial and time domains. This is now possible by exploiting high resolution satellite images and long time series of images and Earth observation products, along with powerful statistical techniques to process them. However, in recent years, the big and heterogeneous data streams acquired by satellite constellations hamper the adoption of advanced machine learning statistical techniques for anomaly detection and anomalous change detection events. The Thesis' main objective is to develop and apply novel and robust detectors, under different framework such as distance, kernel, approximation kernel and probability estimators approaches for the monitoring and protection of areas that are difficult to access using remote sensing images. On the other hand, characterize and implement new detectors of anomalies in several real scenarios such as droughts, wildfires, floods and urbanization from AVIRIS, Sentinel-2, WorldView-2, MODIS, Quickbird and Landsat8 sensors.

Structure and content

The work elaborated in this Thesis is structured in six chapters. First, an introduction to remote sensing for Earth observation and a review of detection methods in ACD-AD-CD settings. In addition, the main methods proposed in this Thesis were proposed based on the kernel's theory and the efficient alternatives to approximate kernel matrices. Also, it is presented a novel method for probability density estimation. The proposed methods based on distance and probability estimation are elaborated in different chapters. These chapters showcase the results that include visual and numerical comparison between methods taking into account performance and robustness. Finally, the thesis is completed with concluding remarks. The Thesis is completed by an annex which includes a compendium of peer-reviewed publications in remote sensing international journals. A prototype of all methods in this thesis is implemented in Matlab software and the code and demos are provided. The outline of each chapter is summarized as follows:

Chapter 2: presents a novel kernelized version of the Cook's distance to address anomalous change detection in remote sensing images. Due to the large computational burden involved in the direct kernelization, and the lack of out-of-sample formulas, this approach introduces and compares both random Fourier features and Nyström

implementations.

Chapter 3: describes the algorithms based on the implicit probability density function (PDF) estimation when the background assumption is Gaussian and Elliptically contoured for ACD. The kernel theory is summarized, including the necessary concepts and properties on reproducing kernel Hilbert spaces (RKHS), present the nonlinear versions based on the kernel theory, and experimentally validate it.

Chapter 4 efficient methods are developed in this chapter to improve the computational cost of nonlinear kernel-based RX anomaly detectors. The wide variety of kernel methods and its theoretical basis were reviewed, and the randomized RX approaches were suggested to develop effective and fast methods.

Chapter 5 proposes an unsupervised method for detecting anomalies and changes in remote sensing images by means of a Gaussianization methodology that allows to estimate multivariate densities accurately. The chapter shows the efficiency of the method in experiments involving both anomaly detection and change detection in different remote sensing image sets.

Chapter 6 summarizes the goals and achievements of this Thesis.

Summary of contributions

This Thesis is a compendium of works based on detecting anomalous changes between pairs of images in remote sensing as well as anomalies when the people are monitoring the Earth cover by means of remote sensing satellite images. Focusing mainly on the development of novel methods based on the kernel theory and theirs approximation when the proposal is represented by distance background. The Thesis includes methods that use directly the Gaussianization transformation of the data to cope with change and anomaly detection approaches under a probabilistic framework. The contributions of the thesis are summarized in Table 1.1.

The Thesis is completed by an annex which includes a compendium of peer-reviewed publications in remote sensing international journals, summarized as follows:

1. *Kernel Anomalous Change Detection for Remote Sensing Imagery*. Padrón-Hidalgo, J. A. and Laparra, V. and Longbotham, N and Camps-Valls, G. IEEE Transactions on Geoscience and Remote Sensing 10, vol 57, pages: 7743-7755, 2019. Journal Impact Factor (5.85). Q1: Electrical and Electronic Engineering. Q1: Remote Sensing.
2. *Efficient Nonlinear RX Anomaly Detectors*. José A. Padrón Hidalgo and Adrián Pérez-Suay and Fatih Nar and Gustau Camps-Valls IEEE Geoscience and Remote Sensing Letters, pages: 1-5, 2020. Journal Impact Factor (3.83). Q1: Electrical and

Table 1.1: Proposed methods in this Thesis: *Methods* corresponds to the implemented detectors, *App*. represents the involved approach of each detector, *Linear* is the assumption of the model description, *Proposed* identifies our proposed methods in this thesis, *Learning* can be either a supervised or unsupervised setting, *Mem*. is the efficiency in computational cost and *Acc*. is the expected accuracy level in remote sensing applications.

Methods	**App.**	**Linear**	**Proposed**	**Learning**	**Mem.**	**Acc.**
Cook	ACD	✓	×	supervised	✓	
Kernel Cook	ACD	×	✓	supervised	×	+
Randomized Cook	ACD	×	✓	supervised	✓	++
Nyström Cook						+++
RX	ACD	✓	×	supervised	✓	
Elliptical-RX						+
Kernel-RX	ACD	×	✓	supervised	×	+
Elliptical Kernel-RX						++
RX	AD	✓	×	supervised	✓	
Kernel-RX	AD	×	✓	supervised	×	+
Subsampling-RX	AD	×	✓	supervised	✓	++
Orthogonal-RX						++
Randomized-RX						++
Nyström-RX						+++
RX	AD-CD	✓	×	unsupervised	✓	++
Kernel-RX	AD-CD	×	✓	unsupervised	×	+
RBIG	AD-CD	×	✓	unsupervised	✓	+++

Electronic Engineering. Q1: Geochemistry and Geophysical.

3. *Efficient Kernel Cook's Distance for Remote Sensing Anomalous Change Detection*. Padrón-Hidalgo, J.A. and Pérez-Suay, A. and Nar, F. and Laparra, V. and Camps-Valls, G. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol 13, pages: 5480 - 5488, 2020. Journal Impact Factor (3.83). Q1: Electrical and Electronic Engineering. Q1: Geographic Physical.
4. *Unsupervised Anomaly and Change Detection with Multivariate Gaussianization*. Padrón-Hidalgo, J. A. and Laparra, V. and Camps-Valls, G. Submitted to IEEE Transactions on Geoscience and Remote Sensing, 2020. Journal Impact Factor (5.85). Q1: Electrical and Electronic Engineering. Q1: Remote Sensing.

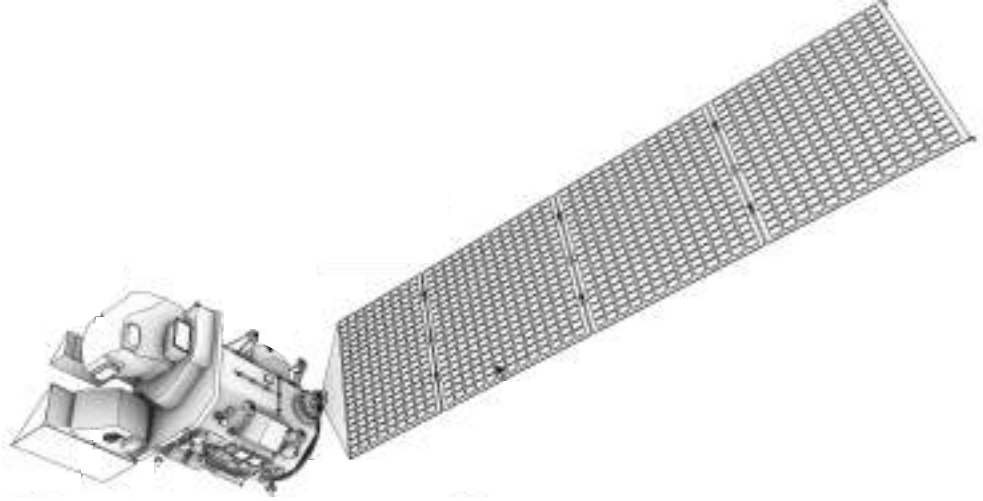
Other related publications in conferences and workshops are listed here too for completeness:

1. *Kernel Anomalous Change Detection*. Jose A. Padrón Hidalgo and Valero Laparra and Gustau Camps-Valls IEEE Young Professionals Conference on Remote Sensing, Aachen, Germany 2018.
2. *Nonlinear Cook Distance for Anomalous Change Detection*. Jose A. Padrón Hi-

dalgo and Adrián Pérez-Suay and Fatih Nar and Gustau Camps-Valls 2018 IEEE International Geoscience and Remote Sensing Symposium, València, Spain 2018

3. *Randomized RX for Target Detection*. Fatih Nar and Adrian Perez-Suay and Jose Antonio Padron and Gustau Camps-Valls 2018 IEEE International Geoscience and Remote Sensing Symposium, València, Spain 2018

All papers are accompanied by supporting material and MATLAB^{TM} source code, datasets, and demos for the sake of reproducibility of the results through the links: (1) Kernel Anomalous Change Detection: http://isp.uv.es/kacd.html, (2) Efficient Nonlinear RX Anomaly Detectors: http://isp.uv.es/code/fastrx.html; (3) The Kernel Cook's Distance: http://isp.uv.es/code/kcook; and (4) Multivariate Gaussianization: https://isp.uv.es/RBIG4AD.html.



2. The Kernel Cook's Distance

Contents

This chapter is partially based on the paper published in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. The authors: José A Padrón Hidalgo, Adrián Pérez-Suay, Fatih Nar, Valero Laparra and Gustau Camps-Valls. Journal Number: 13, pages :5480 - 5488, year: 2020. Journal Impact Factor (3.83).

2.1 Summary

Detecting anomalous changes in remote sensing images is a challenging problem where many approaches and techniques have been presented so far. The standard field is based on multivariate statistics of diagnostic measures which are concerned about the characterization of distributions, detection of anomalies, extreme events and changes. One useful tool to detect multivariate anomalies is the celebrated Cook's distance. Instead of assuming a linear relationship, a novel kernelized version of the Cook's distance is presented to address anomalous change detection in remote sensing images. Due to the large computational burden involved in the direct kernelization, and the lack of out-of-sample formulas, it is introduced and compared both random Fourier features and Nyström implementations of the approximate the solution. The kernel Cook's distance was studied for anomalous change detection in a *chronochrome* scheme, where the anomalousness indicator comes from evaluating the *statistical leverage* of the residuals of regressors between time acquisitions. The performance of all algorithms were illustrated in a representative number of multispectral and very high resolution satellite images involving changes due to droughts, urbanization, wildfires and floods. Very good results and computational efficiency confirm the validity of the approach.

2.2 Kernelized Cook's distance

2.2.1 Notation and the chronochrome approach

Let us define two consecutive d-bands multispectral images in matrix form $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^{n \times d}$ composed of n pixels $\mathbf{x}_i, \mathbf{y}_i \in \mathbb{R}^d$, $i = 1, \ldots, n$. Assume that a set of changes have occurred in between, and that such changes do not alter the image distribution significantly. The 'chronochrome' approach (Schaum & Stocker, 1997) builds on this idea and fits a model to predict the second image $\mathbf{Y}$ from the first one $\mathbf{X}$, and decides that a point is anomalous (i.e. it has changed) if, for instance, the corresponding residual is significantly large. The prediction function $f : \mathbf{x} \to \mathbf{y}$ is learned from the observations. The task is now to assess the significance of the obtained residuals, $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}$, that is to derive a sensible diagnostic measure.

2.2.2 Cook's distance

Cook's distance comes from the definition of *leverage*, which measures how distant are the independent variable values (of a particular observation) from those of the other observations. The highest leveraged points are those observations which could be considered

as extreme or outlying values of the independent variables. Cook's distance measures the effect of removing a given observation. Therefore, the aim is to find out which elements from the sample set are more relevant to the model.

The standard Cook's distance assumes a linear model for prediction of the second image from the first one, i.e. $\hat{\mathbf{Y}} = \tilde{\mathbf{X}}\mathbf{W}$, where $\mathbf{W} \in \mathbb{R}^{(d+1)\times d}$, and $\tilde{\mathbf{X}}$ is the augmented design matrix with a column of ones to account for the bias term, $\tilde{\mathbf{X}} = [\mathbf{X}|\mathbf{1}_n]$. The solution to this least squares problem is given by the Wiener-Hopf normal equations, $\mathbf{W} = (\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}^\top\mathbf{Y}$. The predictions can be expressed as $\hat{\mathbf{Y}} = \tilde{\mathbf{X}}\mathbf{W} = \tilde{\mathbf{X}}(\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}^\top\mathbf{Y} = \mathbf{H}\mathbf{Y}$, where $\mathbf{H}$ is known as the *projection matrix*, and it is defined the *leverage score* of the i-th observation as

$$h_i = \mathbf{x}_i^\top(\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}})^{-1}\mathbf{x}_i. \tag{2.1}$$

Similarly, the i-th element of the residual vector $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = (\mathbf{I} - \mathbf{H})\,\mathbf{y}$ is denoted as e_i. The Cook's distance D_i for observation $\mathbf{x}_i$, $i = 1,\ldots,n$, is defined as the sum of all the changes in the regression model when the i-th observation is deleted:

$$D_i = \frac{\sum_{j=1}^{n}\left(\hat{\mathbf{y}}_j - \hat{\mathbf{y}}_{j\setminus i}\right)^2}{d\ \mathrm{MSE}^2}, \tag{2.2}$$

where $\hat{\mathbf{y}}_j$ means to predict the j-th sample through the model trained with all the samples and $\hat{\mathbf{y}}_{j\setminus i}$ is the fitted response value obtained when i is excluded, and MSE is the mean-square error of the regression model with all samples, i.e. $MSE = \frac{1}{N}\sum_{j=1}^{n}\left(\hat{\mathbf{y}}_j - \mathbf{y}_j\right)^2$. Cook's distance can be equivalently expressed using the leverage

$$D_i = \frac{e_i^2 h_i}{d\ \mathrm{MSE}^2(1-h_i)^2}. \tag{2.3}$$

Cook showed that this estimation can be obtained using incremental rank-one updates of covariances, without even needing to re-compute each model when the i-th sample is removed (Cook, 1977).

2.2.3 Kernel Theory

This section includes a brief introduction to kernel methods. After setting the scenario and fixing the most common notation, the main properties of kernel methods are provided. Also, pay attention to kernel methods development by means of particular properties drawn from linear algebra and functional analysis.

Kernel methods measure similarities between samples mapped into a Hilbert space $\mathcal{H}$

of higher dimensionality. The dot products therein are not estimated explicitly, but through a reproducing kernel function that approximates it. Actually, kernel methods do not require to have access to the feature map to $\mathcal{H}$, neither to compute the data coordinates in $\mathcal{H}$ to estimate similarities, which can be done implicitly via reproducing kernel functions. Given a dataset with input feature vector $\mathbf{x} \in \mathcal{X}$, the feature mapping can be defined by $\phi : \mathcal{X} \to \mathcal{H}$, hence $\mathbf{x} \mapsto \phi(x)$. Therefore, the similarity between the elements in $\mathcal{H}$ can now be measured using its associated dot product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$. A kernel function computes the similarity between inputs such that $(\mathbf{x}, \mathbf{x}') \to K(\mathbf{x}, \mathbf{x}')$ and the function satisfies:

$$K(\mathbf{x}, \mathbf{x}') = \langle \phi(\mathbf{x}) \phi(\mathbf{x}') \rangle_{\mathcal{H}}. \tag{2.4}$$

The mapping ϕ is called *feature map* and the space $\mathcal{H}$ is its corresponding *feature space*. In addition, 2.4 is also known in the machine learning literature as the kernel trick which states that all dot products in $\mathcal{H}$ can be implicitly computed by simply using a kernel function defined on the input data. If one have access to a dataset of n examples, $\mathbf{x}_i$, $i = 1, \ldots, n$, then a function will denoted the similarity as the set of similarities $f(\cdot) = [K(\mathbf{x}_1, \cdot), \ldots, K(\mathbf{x}_n, \cdot)]$, and will denote $\mathbf{K}$ the kernel matrix that contains all similarities among the n data points, which has entries $[\mathbf{K}]_{ij} = [K(\mathbf{x}_i, \mathbf{x}_j)]$.

Kernel Ridge Regression

Now that the basic theory that underlies kernel spaces is covered, more practical issues can be addressed. Let us derive the first kernel method that will become a core in the thesis: the kernel ridge regression (KRR), which will be used in our proposed Kernel Cook's distance later. KRR is a nonlinear version for fit a linear model in Hilbert spaces, so the prediction model is given by $\hat{\mathbf{Y}} = \mathbf{\Phi} \mathbf{W}_{\mathcal{H}}$. The weights $\mathbf{W}_{\mathcal{H}}$ (including a bias term for simplicity) are calculated using the regularized loss function:

$$L = \|\mathbf{Y} - \mathbf{\Phi} \mathbf{W}_{\mathcal{H}}\|^2 + \lambda \|\mathbf{W}_{\mathcal{H}}\|^2 .$$

The representer's theorem states that one can express the solution matrix $\mathbf{W}_{\mathcal{H}}$ defined in $\mathcal{H}$ as a linear combination of mapped samples in the RKHS, hence $\mathbf{W}_{\mathcal{H}} = \mathbf{\Phi}^\top \boldsymbol{\alpha}$. Now, following the standard least squares solution, the primal solution can be described as

$$\mathbf{W}_{\mathcal{H}} = (\mathbf{\Phi}^\top \mathbf{\Phi} + \lambda \mathbf{I})^{-1} \mathbf{\Phi}^\top \mathbf{Y}$$

where $\mathbf{\Phi} \in \mathbb{R}^{n \times D_{\mathcal{H}}}$. The next step is replacing the inner product by a kernel function using

the dual solution:

$$\boldsymbol{\alpha} = (\boldsymbol{\Phi}\boldsymbol{\Phi}^\top + \lambda \mathbf{I})^{-1}\mathbf{Y} = (\mathbf{K} + \lambda \mathbf{I})^{-1}\mathbf{Y}.$$

Then, once obtained the dual solution, it is easy to calculate the prediction for new m data $\mathbf{X}_* \in \mathbb{R}^{m\times d}$, and their mappings $\boldsymbol{\Phi}_*$, as follows:

$$\hat{\mathbf{Y}} = \boldsymbol{\Phi}_* \mathbf{W}_{\mathcal{H}} = \boldsymbol{\Phi}_* \boldsymbol{\Phi}^\top \boldsymbol{\alpha} = \boldsymbol{\Phi}_* \boldsymbol{\Phi}^\top (\boldsymbol{\Phi}_* \boldsymbol{\Phi}^\top + \lambda \mathbf{I})^{-1}\mathbf{Y}$$

Finally, by applying the kernel trick, one can replace the inner product by the kernel evaluations (similarities) between the corresponding training or test samples:

$$\hat{\mathbf{Y}} = \mathbf{K}_{*:}(\mathbf{K} + \lambda \mathbf{I})^{-1}\mathbf{Y},$$

where $\mathbf{K}_{mn} \in \mathbb{R}^{m\times n}$ and $\mathbf{K}_{nn} \in \mathbb{R}^{n\times n}$. This formulation help us to understand in an easy manner the kernel Cook's distance approximation in the next section.

2.2.4 Kernel Cook's distance

The kernel Cook's distance (KC) can be easily derived by departing from Eq. (2.3). For that, both the errors and the leverage scores must be calculated as a function of the input data only. Let us first recall the KRR prediction formula, $\hat{\mathbf{y}} = \mathbf{K}(\mathbf{K} + \lambda \mathbf{I})^{-1}\mathbf{y}$, where λ is a regularization parameter, and $\mathbf{K}$ is the kernel matrix. The residuals are thus $\mathbf{e} = (\mathbf{I} - \mathbf{H}^{\mathcal{H}})\mathbf{y}$, where the (kernel) projection matrix $\mathbf{H}^{\mathcal{H}} = \mathbf{K}(\mathbf{K} + \lambda \mathbf{I})^{-1}$, and the (kernel) leveraging scores become

$$h_i^{\mathcal{H}} = \operatorname{diag}(\mathbf{H}^{\mathcal{H}}), \quad i = 1, \ldots, n \tag{2.5}$$

From here, one can readily compute $e_i^{\mathcal{H}}$ and the kernel Cook's distance as:

$$D_i^{\mathcal{H}} = \frac{(e_i^{\mathcal{H}})^2}{d\ \mathrm{MSE}^2} \frac{h_i^{\mathcal{H}}}{(1 - h_i^{\mathcal{H}})^2}. \tag{2.6}$$

Note that the the inversion of a large $\mathbf{K}$ matrix in $\mathbf{H}^{\mathcal{H}}$ has a cost of cubic time complexity and quadratic space (memory) complexity. One could think of computing the leverage scores using a singular value decomposition (SVD), but the exact computation is as costly as solving the original problem since the cost is also cubic. Unlike the linear case, the recursive solution of (2.6) is cumbersome and one has to recompute each model after

sample deletion, thus involving a cascade of costly inverse operations.

2.3 Efficiency in Kernel Cook

In this section, both random Fourier features and Nyström approximation of the leverage scores and the errors for Cook's distance approximation will exploited.

2.3.1 Randomized Cook's distance

Let us first approximate the kernel matrix with random Fourier features (Rahimi & Recht, 2007). Formally, a linear regression model expressed on data explicitly projected onto q random Fourier features is used. Let us define a feature map $\mathbf{z}(\mathbf{x}) : \mathbb{R}^d \to \mathbb{C}^q$, *explicitly* constructed as $\mathbf{z}(\mathbf{x}) := [\exp(\mathbb{i}\mathbf{w}_1^\top \mathbf{x}), \ldots, \exp(\mathbb{i}\mathbf{w}_q^\top \mathbf{x})]^\top$, where $\mathbb{i} = \sqrt{-1}$, and $\mathbf{w}_q \in \mathbb{R}^d$ is randomly sampled from a data-independent distribution (Rahimi & Recht, 2007). The prediction model is now defined as $\hat{\mathbf{Y}} = \Re\{\mathbf{ZW}\}$, where $\mathbf{Z} = [\mathbf{z}_1 \cdots \mathbf{z}_n]^\top \in \mathbb{R}^{n\times q}$, with the weight matrix $\mathbf{W} \in \mathbb{R}^{q\times d}$. The *randomized leverage* of a particular sample is now expressed

$$h_i^R = \Re\{\mathbf{z}(\mathbf{x}_i)^\top (\mathbf{Z}^\top \mathbf{Z} + \lambda \mathbf{I})^{-1} \mathbf{z}(\mathbf{x}_i)\}, \tag{2.7}$$

which is then plugged into (2.3) owing to the linearity of the model where $\mathbf{e}^R = (\mathbf{I} - \mathbf{H}^R)\mathbf{y}$ and then leads to

$$D_i^R = \frac{(e_i^R)^2}{d\ \mathrm{MSE}^2} \frac{h_i^R}{(1-h_i^R)^2}. \tag{2.8}$$

This allows to control the memory and computational complexity explicitly through q, as one has to store matrices of $n \times q$ and invert matrices of size $q \times q$ only. It is worth noting that, in practice, a low number of random Fourier features are needed, $q \ll n$. This is not only beneficial in computation time and memory savings but also has a regularization effect in the solution.

2.3.2 Nyström Cook's Distance

The Nyström method selects a small set of $r \ll n$ samples to make a low-rank approximation of an $n \times n$ kernel matrix $\mathbf{K} \approx \mathbf{K}_{rn}^\top \mathbf{K}_{rr}^{-1} \mathbf{K}_{rn}$ (Williams & Seeger, 2001), where $\mathbf{K}_{rn} \in \mathbb{R}^{r\times n}$ contains the kernel similarities between $\hat{\mathbf{X}} \in \mathbb{R}^{r\times d}$ and $\mathbf{X} \in \mathbb{R}^{n\times d}$, and $\mathbf{K}_{rr} \in \mathbb{R}^{r\times r}$ is a kernel matrix containing data similarities between the points in $\hat{\mathbf{X}}$. By exploiting the

Table 2.1: Space and time complexity for all methods: **T** is transformation of image into a nonlinear space, **C** is for covariance/kernel matrix, **W** is for regression weight, **L** is for leverage, **ACD** is the Cook's distance, and $\mathcal{O}(.)$ is the overall complexity.

Method	**T**	**C**	$\mathbf{C^{-1}}$	**W**	**L**	**ACD**	$\mathcal{O}(.)$
Space							
L-Cook	–	d^2	d^2	d^2	n	n	$\mathcal{O}(nd)$
R-Cook	nq	q^2	q^2	q^2	n	n	$\mathcal{O}(nq)$
N-Cook	n^2	r^2	r^2	–	n	n	$\mathcal{O}(n^2)$
K-Cook	n^2	n^2	n^2	–	n	n	$\mathcal{O}(n^2)$
Time							
L-Cook	–	nd^2	d^3	nd^2	nd^2	nd^2	$\mathcal{O}(nd^2)$
R-Cook	nqd	nq^2	q^3	nq^2	nq^2	nq^2	$\mathcal{O}(nq^2)$
N-Cook	n^2d	nr^2	r^3	–	n^2r	n^2d	$\mathcal{O}(n^2r)$
K-Cook	n^2d	n^3	n^3	–	n^3	n^2d	$\mathcal{O}(n^3)$

Nyström method in the Woodbury-Morrison the fallowing formula is obtained:

$$(\mathbf{K}+\lambda\mathbf{I})^{-1} = \lambda^{-1}(\mathbf{I}-\mathbf{K}_{nr}(\lambda\mathbf{K}_{rr}+\mathbf{K}_{nr}^{\top}\mathbf{K}_{nr})^{-1}\mathbf{K}_{nr}^{\top}), \tag{2.9}$$

and now defining $\mathbf{Q} = \lambda\mathbf{K}_{rr}+\mathbf{K}_{nr}^{\top}\mathbf{K}_{nr}$, the projection matrix approximation is defined as:

$$\mathbf{H}^N = \lambda^{-1}\mathbf{K}(\mathbf{I}-\mathbf{K}_{nr}\mathbf{Q}^{-1}\mathbf{K}_{nr}^{\top}), \tag{2.10}$$

with Nyström leverage scores

$$h_i^N = \text{diag}(\mathbf{H}^n), \tag{2.11}$$

and $\mathbf{e}^N = (\mathbf{I}-\mathbf{H}^N)\mathbf{y}$, thus the Nyström Cook's distance becomes:

$$D_i^N = \frac{(e_i^N)^2}{d\ \text{MSE}^2}\frac{h_i^N}{(1-h_i^N)^2}. \tag{2.12}$$

2.3.3 Memory and computational cost

Space (memory) and time (computational) efficiency of the linear and nonlinear versions are presented in Table 2.1. In this approach, the linear version is named as L-Cook while the nonlinear versions are named Randomized Cook (R-Cook), Nyström Cook (N-Cook), and Kernel Cook (K-Cook). Note that, d is the spectral dimension and it is around 10 for multispectral images and around 100 for hyperspectral images. Although q and r can have

similar values, generally $q < r$. Since large images are used, n is much larger than r, q, and d. Therefore, in general $d < q < r \ll n$.

As it can be seen in Table 2.1, the L-Cook method provides superior space and time efficiency. However, the L-Cook method is only limited to rare linear scenarios where the real-world nonlinear transformation between multi-temporal images are formed due to various reasons. However, space and time complexity of the K-Cook method is proportional to the number of pixels in the image, respectively quadratic in space and cubic in time. Thus, the use of the K-Cook method is not feasible for large images, which is the common scenario nowadays. Note that, for the N-Cook method, kernel matrix $\mathbf{K}$ is still used in (2.10) but there is no inversion operation on it. Therefore, N-Cook has same space complexity with K-Cook method since it needs to store kernel matrix $\mathbf{K}$. But time complexity of N-Cook is still superior to the K-Cook method since only an $r \times r$ matrix is inverted.

2.4 Experimental Results

This section analyzes the performance of the proposed linear and nonlinear Cook's distance methods for anomalous change detection. In order to test the robustness of the proposed methods, tests were performed in both simulated and real scenes with changes. The detection performance of the methods were evaluated quantitatively through the Area Under the Curve (AUC) of the Receiver Operating Characteristic (ROC) and qualitatively by inspection of the detection maps. Two experiments have been performed with different complexities of difficulty while controlling the analyzed changes. The first experiment is designed over a real scenario and synthetic changes. The second set of experiments deals with both real scenes and natural changes related to floods, fires and urbanization. In order to ease the reproducibility, the MATLAB implementations of the methods have been provided. Moreover were made available a database with the labeled images used in the second experiment in http://isp.uv.es/code/kcook.

2.4.1 Experiment 1: Real Scene with Simulated Changes

The aim of this experiment is to show and analyze the performance of the proposed methods when the change between images is nonlinearly distributed. In this example, one can analyze how nonlinear methods fit the regression model to the data well and how they detect the influential points in the Cook's distance approach. The experiment involves representing a nonlinear relation between two images in order to demonstrate the limitations of the linear algorithms in this situation.

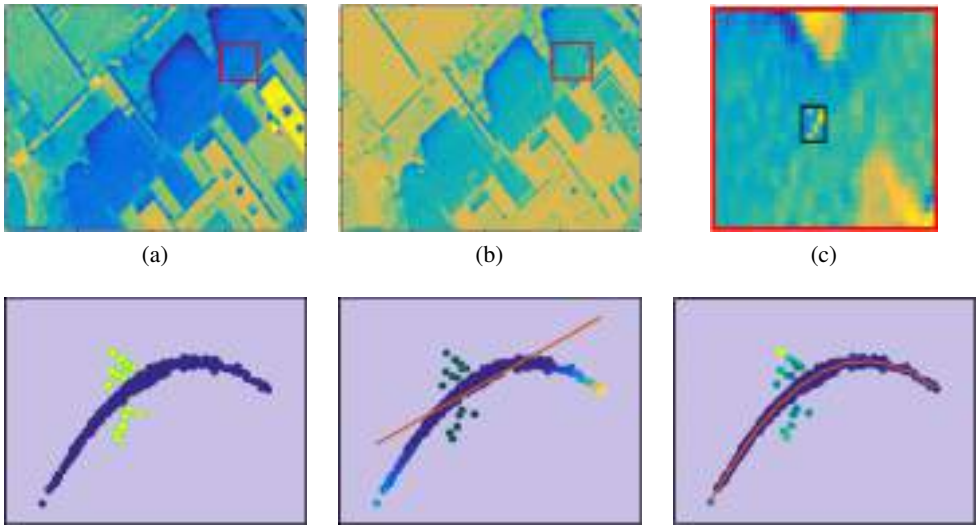


Figure 2.1: (a) Image (R band) at time t_1 and the region of interest (red box). (b) Image (R band) at time t_2 and the region of interest (red box). The background color distortion was applied and square patches of 4×4 were added over t_2 simulating the anomalies, (c) region of interest (red box in t_2) and the corresponding label is surrounded and highlighted in black, (d) scatter plots between t_1 and t_2 pixels in R band, blue dots represent the non-change class and the yellow dots correspond to change class. Panel (e) shows how mis-specification of the linear regression model cannot detect anomalies, while a nonlinear Cook's distance can do in (f). In both (e) and (f) the dots color specify how much anomalous the point is for the model (blue less, yellow more).

Figure 2.1 (a)-(b) show an aerial scene taken over the Image Processing Laboratory (IPL) from Google Earth in the R band. Figure 2.1 (a) represent the image at time t_1 (no change class), while Figure 2.1 (b) represent the image at time t_2 (change class). All the values of the second image (t_2) were modified by applying a soft nonlinear function (an inverted parabola) to simulate non anomalous changes. In order to introduce the anomalous changes, square patches of 4×4 pixels randomly selected were interchanged.

Since kernel Cook's distance is computationally very demanding, a portion of the full image have been selected in order to have a comparison of all proposed methods together. In particular, the region of interest is shown in Fig. 2.1 (c) and marked in a red box in Fig. 2.1 (b), the anomalies are highlighted in a black rectangle and the anomalous class represent the 0.016%. Figure 2.1 (d) represents the scatter of original image x-axis against transformed image y-axis, the points in yellow color are the change pixels but the points in blue color ideally would not be detected as an anomalous change pixels. Figure 2.1 (e) illustrates how a linear model does not fit the distribution well and the inferred values lead to False Positives errors (in the tails) and True Negatives errors (green color). Figure 2.1 (f) shows how a nonlinear model over distribution fits well and both avoid the False Positives

and detects the changed pixels in the images. These results are confirmed visually through the prediction maps in Fig. 2.2, where the kernel Cook's distance excels in detection.

(a) (b) (c)

Figure 2.2: (a) represent the prediction map (labels), (b) display the change prediction map detected by the linear method and (c) the change prediction map detected by the nonlinear Cook's distance.

On the other hand, Table 2.2 showcase how efficient the proposed efficient methods can be, achieving better values of AUC compared to the kernel one in less time. Therefore, because of the huge computational cost involved in its calculation, one cannot use it in standard images (even as small as the one in Fig. 2.1), so efficient algorithms for computing Cook's distances in nonlinear kernel settings are strictly necessary.

Table 2.2: Area under the curve (AUC) and their respective Time values (in seconds) per method.

Methods	L-Cook	R-Cook	N-Cook	K-Cook
AUC	0.55	0.93	0.93	0.92
Time	0.01	0.03	2.64	6.32

2.4.2 Experiment 2: Real and Natural Changes

In this section experiments in several real satellite images are reported. The aim is to detect changes that can be found naturally in a real environment. The dataset is composed of five different scenes with natural changes including urbanization, wildfires, droughts and flooding.

Data collection

Pairs of multispectral images acquired at different times over the same location were collected. The images were selected in such a way that a noticeable change happened between the two acquisition times. Was photo-interpreted and manually labeled all the image pixels affected by a change of interest. This step is critical and delicate since one could fall into many false alarms due to, for instance, shadows, illumination changes or natural changes in the vegetation. All images contain changes of a different nature, which allows us to study how the different Cook's distance algorithms perform in a diversity of

realistic scenarios.

A brief summary of the images and change events follows. Argentina dataset represent an area burned between the months of July and August 2016. Denver Region Urbanized Project Area describes the stereo-compiled building roofprints feature of Denver Regional Council of Government (DRCOG). Texas wildfire dataset is composed by a set of four images acquired by different sensors over Bastrop County, Texas (USA), and is composed by a Landsat 5 TM as the pre-event image and a Landsat 5 TM plus an EO-1 ALI and a Landsat 8 as post-event images. This phenomenon is considered the most destructive wildland-urban interface wildfire in Texas history. The Arizona dataset corresponds to the decline of Lake Powell in USA. The first image was taken by Landsat-5 and shows its highest water level. The second was taken by Landsat-8 following a period of drought that began in 2000. When the water volume was measured five months later, it was less than half of the maximum lake capacity. The Australia dataset shows the natural floods caused by Cyclone Debbie in Australia 2017. Storm damage resulted from both the high winds associated with the cyclone, and the very heavy rain that produced major riverine floods. Table 2.3 gives some descriptors of the images in the database, while Fig. 2.3 shows the RGB composites of the pairs of images and the corresponding reference map.

Numerical comparison

The hyperparameters using 1000 randomly selected pixels were selected for cross-validation. Each method implies a different set of parameters. For both the randomized and Nyström methods have been cross-validated the r and q parameters by exploring values between 1 and 400, particularly $r,q \in \{1,5,10,25,50,100,200,300,400\}$.In this approach, the standard Radial Basis Function (RBF) kernel function was used, $K(\mathbf{x},\mathbf{x}') = \exp(-\|\mathbf{x}-\mathbf{x}'\|^2/(2\sigma^2))$. The RBF kernel shows good theoretical properties (universal kernel, smoothness and robustness), convenience (only the lengthscale parameter σ needs to be tuned) and good performance in practice. The RBF kernel is used to perform kernel regression, which incorporates a regularization parameter λ. The σ and λ parameters were searched

Table 2.3: Images attributes used in the experimentation dataset.

Images	Sensor	Size	Bands	Resolution (m)
Argentina	Sentinel-2	381 x 500	12	10-60
Denver	Quickbird	101 x 101	4	0.6-2.4
Arizona	Cross-Sensor	201 x 201	7	30
Texas	Cross-Sensor	301 x 201	7	30
Australia	Sentinel-2	201 x 501	12	10-60

using a logarithmic grid between 10^{-4} and 10^{20}.

The hyper-parameters of the different methods were optimized to maximize the cross-validation AUC. The ROCs and Precision-Recall curves were compared in terms of AUCs for all methods and images in Fig. 2.4. In general, all methods can cope with the large dimension of the images, and can provide reasonable results, AUC>0.70, see Table 2.4.

Table 2.4: Area under the curve (AUC) per method and scene. The best results are bold faced.

Methods	Argentina	Denver	Arizona	Texas	Australia
L-Cook	0.91	0.83	0.75	0.91	0.69
R-Cook	0.93	0.87	0.77	0.92	0.79
N-Cook	**0.93**	**0.96**	**0.99**	**0.97**	**0.94**

The nonlinear versions (randomized and Nyström approximations) improve the results of the linear Cook's distance, revealing nonlinear changes in all scenes, yet differences are minor for the Texas scene. The Nyström Cook's distance achieves consistently the best results in all the scenarios, and false or positive rates regimes. A average gain of +15.6% over the linear approach, and of +11.8% over the randomized approach, along with the computational efficiency justify the adoption of this approach. The double logarithmic plot aims to better appreciate the differences in very low false positive rates regimes. Also, precision and recall are an understanding and measure of relevance. Here it becomes clear that the Nyström approach excels in all images.

For each experiments 1000 runs were made for testing the significance of the methods based on the ROC profiles. The mean value of the experimental runs is plotted with the standard deviation of each detection algorithm represented by the shaded region in Fig. 2.5. Also, a boxplot is showed in the same figure to illustrate the standard deviation of each methods with a better precision. As seen in Fig. 2.5, N-Cook has always superior or equivalent performance compared to L-Cook and R-Cook, i.e. higher detection rate and lower false alarm rate, and higher AUC value and lower standard deviation.

Visual comparison

A visual comparison of the results is given in Fig. 2.3. Differences between the L-Cook and the R-Cook are not visually significant either. In general, N-Cook yields clear and sharper detection maps (last column), especially in large spatial structures (see e.g. roofs in Denver, lake in Arizona) but also exhibits a much lower false alarm rate (see e.g. a less amount of spurious detections in Texas wildfires). This is sometimes compensated with sensitivity to subtle reflectance changes and misclassified pixels in Australia due to imperfect labeling of pixels. This is why this problem is so difficult to solve in an automatic way.

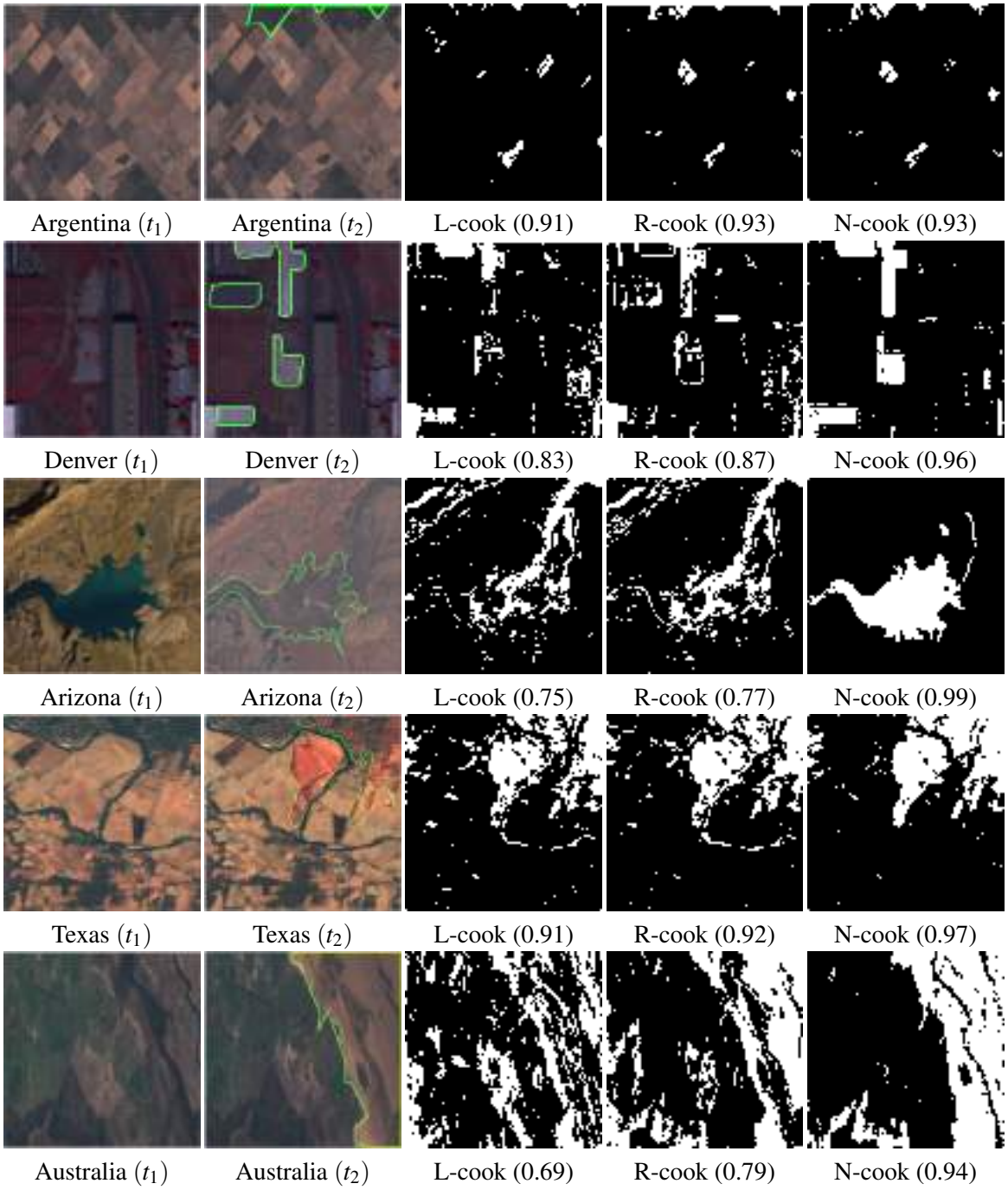


Figure 2.3: RGB composite images and predictions maps. First row: represent an area burned between the months of July and August 2016 (Argentina), anomalous samples represent 2.7%. Second row: urbanization area over Denver city correspond to roofprints (extension of anomalous pixels represents the 11.5% of the image). Third row: decline of the Lake Powell in Arizona, USA (16.35%). Fourth row: the most destructive wildland-urban interface wildfire in Texas history (19.5%). Last row: natural floods caused by Cyclone Debbie in Australia (34%). First column: images without changes, first time of acquisition (t_1). Second column: images with the anomalous changes and their corresponding labels are surrounded and highlighted with green color, second time of acquisition (t_2). Third column: prediction map of linear method. Fourth column: prediction map of random Fourier features method. Last column: prediction map of Nyström approximation method. AUC value in parentheses.

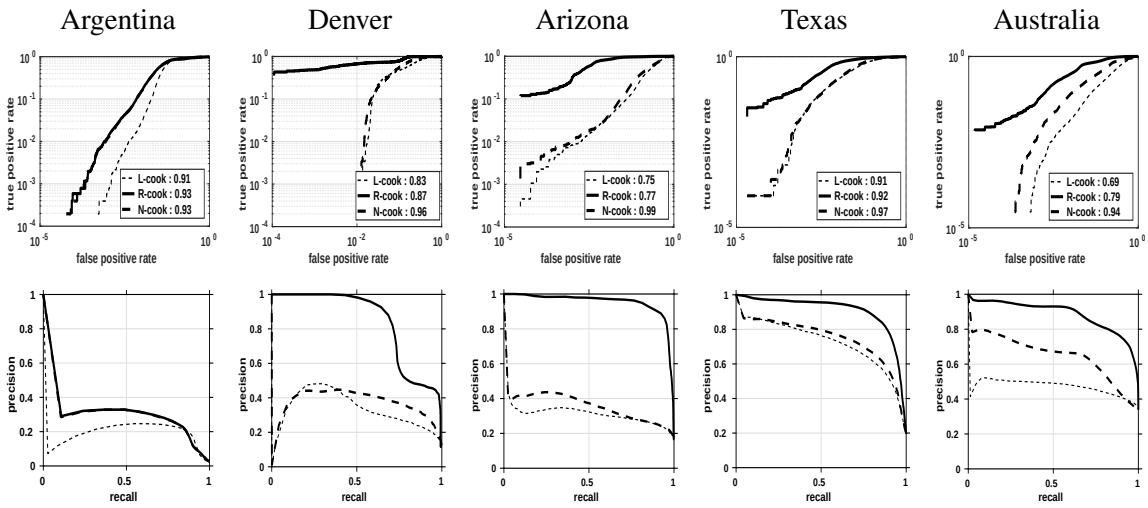


Figure 2.4: ROC curves and Precision-Recall for all images by columns. First row showcase the ROC curves in logarithmic scale. Numbers in legend display the AUC values for each method. Second row showcase the precision-recall following the ROC curves legend.

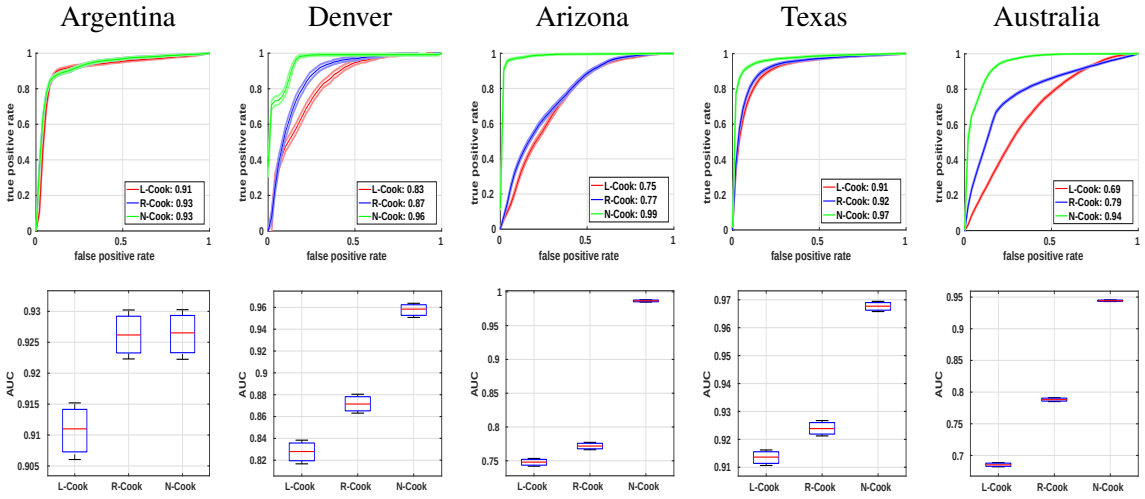
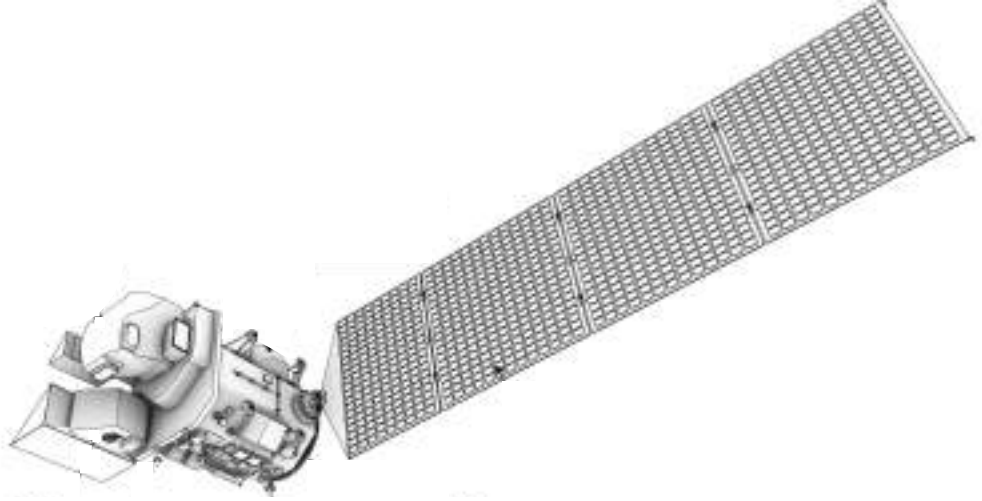


Figure 2.5: Bootstrap experiment. Top row correspond to the ROC curves taken account the mean value of the 1000 iterations. The standard deviation of each approach is illustrated by the shaded region. In the bottom row, AUC values and standard deviation for each method are shown as boxplot.

2.5 Specific contributions

This chapter focuses on introducing a family of efficient nonlinear ACD algorithms based on the Cook's distance setting. The theory of reproducing kernels was used, and proposed several efficient approximation methods following the standard linear approach. The kernel Cook detector was developed and improved using efficient and fast techniques based on feature maps and low-rank approximations wich allows to find influential points (anomalies) in the chronochrome setting. In this chapter, an exhaustive statistical experimentation over simulated and real data scenes based on the study of the ROC and Precision Recall curves have been developed to achieve maximum performance in the AUC measure. Among all methods, the Nyström approximation achieves the best results and yields a more efficient and accurate non-linear method to be applied in practice.



3. KERNEL ANOMALOUS CHANGE DETECTION

Contents

This chapter is partially based on the paper published in IEEE Transactions on Geoscience and Remote Sensing. The authors: José A Padrón Hidalgo, Nathan Longbotham, Valero Laparra and Gustau Camps-Valls. Journal Number: 10, Volume: 57, pages :7743-7755, year: 2019. Journal Impact Factor (5.85).

3.1 Summary

Anomalous change detection is an important problem in remote sensing image processing. Detecting not only pervasive but anomalous or extreme changes has many applications for which methodologies are available. This chapter introduces a nonlinear extension of a full family of anomalous change detectors. In particular, it is focused on algorithms that utilize Gaussian and elliptically contoured distribution and extend them to their nonlinear counterparts based on the theory of reproducing kernels Hilbert space. The performance of the introduced kernel methods are illustrated in both pervasive and anomalous change detection problems with real and simulated changes in multi and hyperspectral imagery with different resolutions (AVIRIS, Sentinel-2, WorldView-2, Quickbird). A wide range of situations are studied in real examples, including droughts, wildfires, and urbanization. Excellent performance in terms of detection accuracy compared to linear formulations is achieved, resulting in improved detection accuracy and reduced false alarm rates. Results also reveal that the elliptically-contoured assumption may be still valid in Hilbert spaces.

3.2 Statistical view of anomalous change detection problem

Anomalies can be loosely defined as rare items, i.e. with low probability to occur (Yuan et al., 2016a,b). Also, it's sometimes referred to as outlier, novelty or extreme detection. An anomalous change is thus a rare, unexpected, change between two consecutive observations. This chapter is focused on finding samples that can be interpreted as anomalous changes between two multidimensional images. This calls for studying and characterizing differences between multivariate distributions, and in particular those features that account for the anomalous changes. In (Theiler et al., 2010) a framework to define different anomalous change detectors based on probability distributions was formalized.

Given two images (X and Y) one can treat their pixel values ($\mathbf{x}_i, \mathbf{y}_i$, with $i = 1, \ldots, N$, where N is the number of pixels) as random variables, with probability distributions $\mathbf{x} \sim \mathbb{P}_X$ and $\mathbf{y} \sim \mathbb{P}_Y$, respectively. These distributions can be used to assess how anomalous is each pixel inside each particular image. On the other hand, let us indicate the joint distribution as $[\mathbf{x}, \mathbf{y}] \sim \mathbb{P}_{X,Y}$, which accounts for how probable particular joint pixel values are, or equivalently to characterize how anomalous a particular change is. For example, if a pixel value changes from $\mathbf{x}_i$ to $\mathbf{y}_i$ and this change has a high probability of occur, it will be classified as a regular change and will not be detected as an anomaly, even if the magnitude change between $\mathbf{x}_i$ and $\mathbf{y}_i$ is highly striking.

The idea is to combine both informations to spot only the changes that are not regular.

Given two pixels $\mathbf{x}_i, \mathbf{y}_i \in \mathbb{R}^d$ from the same spatial location i but each one from one image, the general formula to compute the amount of anomalousness of a change is

$$\widehat{\mathcal{A}}_{[X,Y]}(\mathbf{x}_i, \mathbf{y}_i) = \frac{\mathbb{P}_X(\mathbf{x}_i)\mathbb{P}_Y(\mathbf{y}_i)}{\mathbb{P}_{[X,Y]}(\mathbf{x}_i, \mathbf{y}_i)}. \tag{3.1}$$

A sample is detected as anomalous change when it is anomalous with respect to the joint distribution but not anomalous with respect to the distributions of each isolated image. Here, all the three distributions are used, however different combinations can be used, as shown below.

Instead of using directly Eq. 3.1, it is usual to apply it taking logarithms (Theiler et al., 2010), $\mathcal{A}_{[X,Y]}(\mathbf{x}_i, \mathbf{y}_i) = \log(\widehat{\mathcal{A}}_{[X,Y]}(\mathbf{x}_i, \mathbf{y}_i))$. This can be interpreted in information theoretic terms by noting the relation between probability and information. Elaborating on Shannon's information (Shannon, 2001) may be described as:

$$\mathcal{A}_{[X,Y]}(\mathbf{x}_i, \mathbf{y}_i) = I_{[X,Y]}([\mathbf{x}_i, \mathbf{y}_i]) - I_X(\mathbf{x}_i) - I_Y(\mathbf{y}_i),$$

where $I_A(\mathbf{b})$ is the amount of information in Shannon's terms the sample $\mathbf{b}$ provides assuming it follows the distribution $\mathbb{P}_A$. In this terms a sample will be interpreted as an anomalous change if the information obtained by observing the sample in both images simultaneously is big with respect to the information obtained by observing it in each isolated image.

3.3 Linear ACD algorithms

Assuming that all three distributions follow a *multivariate Gaussian* one can express the formula only in terms of covariance matrices. The amount of *anomalousness* is given by:

$$\mathcal{A}_{\mathcal{G}}(\mathbf{x}_i, \mathbf{y}_i) = \xi(\mathbf{z}_i) - \beta_x \xi(\mathbf{x}_i) - \beta_y \xi(\mathbf{y}_i), \tag{3.2}$$

where $\xi(\mathbf{a}) = \mathbf{a}^\top \mathbf{C}_a^{-1} \mathbf{a}$, $\mathbf{C}_a$ is the estimated covariance matrix with the available data, and being $\mathbf{z} = [\mathbf{x}, \mathbf{y}] \in \mathbb{R}^{2d}$. The value of $\beta_x, \beta_y \in \{0, 1\}$ parameters defines which distributions are taken into account to define our anomaly. The different combinations give rise to different anomaly detectors (see Table 3.1). These methods and some variants have been widely used in many hyperspectral image analysis settings because of its simplicity and generally good performance (Chang & Chiang, 2002; Kwon et al., 2003; Reed & Yu, 1990a).

Table 3.1: A family of ACD algorithms.

ACD algorithm	β_x	β_y
RX	0	0
Chronocrome $\mathbf{y}\|\mathbf{x}$	0	1
Chronocrome $\mathbf{x}\|\mathbf{y}$	1	0
Hyperbolic ACD	1	1

However, these methods are hampered by a fundamental problem: the (typically strong) assumption of Gaussianity that is implicit in the formulation. Accommodating other data distributions may not be easy in general. Theiler et al. (Theiler et al., 2010) introduced alternative ACD to cope with elliptically-contoured distributions (Cambanis et al., 1981): roughly speaking, the idea is to model the data using an *elliptically-contoured (EC) distribution*. EC distributions are particularly convenient in the case of images (Lyu & Simoncelli, 2009). In particular the formulation introduced in (Theiler et al., 2010) uses the multivariate Student's t-distribution, giving rise to the following formula for computing the amount of *EC anomalousness*:

$$\begin{aligned} \mathcal{A}_{\mathrm{EC}}(\mathbf{x}_i,\mathbf{y}_i) &= (2d+\nu)\log\left(1+\frac{\xi(\mathbf{z}_i)}{\nu}\right) \\ &- \beta_x(d+\nu)\log\left(1+\frac{\xi(\mathbf{x}_i)}{\nu}\right) \\ &- \beta_y(d+\nu)\log\left(1+\frac{\xi(\mathbf{y}_i)}{\nu}\right), \end{aligned} \tag{3.3}$$

where ν controls the shape of the Student's t-distribution: for $\nu \to \infty$ the solution approximates the Gaussian and for $\nu \to 0$ it diverges.

An interesting particular case is the RX algorithm which brings to the same result for the Gaussian and the elliptical case (independently of the ν value). All extra operations applied by the EC formulation with regard to the Gaussian version are increasing monotonic functions which do not change the ordering of the values. Therefore, although the values of anomalousness are different (i.e. $\mathcal{A}_{\mathcal{G}}(\mathbf{x}_i,\mathbf{y}_i) \neq \mathcal{A}_{\mathrm{EC}}(\mathbf{x}_i,\mathbf{y}_i)$), the values are sorted in the same way which makes the detection curves equal too. The same effect happen between the RX methods based on kernels proposed in the next section.

Figure 3.1 shows an example of the distributions involved in the anomalous change detection setting. In order to be able to visualize the distributions, a simple situation is shown in which each image contains just one band. In particular, the distribution provided correspond to the band 9 of a Sentinel-2 image over Australia, see table 4.2. The results are shown for the distribution of the data estimated using histograms, and when

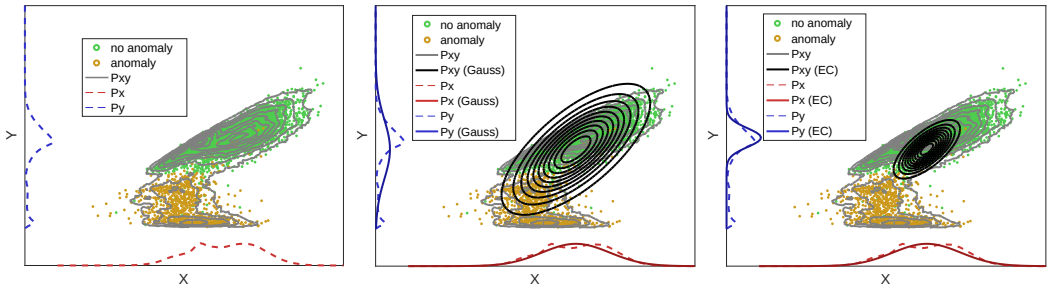


Figure 3.1: Description of probabilistic framework for ACD. From left to right: the original data, Gaussian model, and Elliptically Contoured model. See text for details.

assuming Gaussian or EC distributions. Note that the estimation of the distribution based on histograms is only feasible in the low dimensional (i.e. 2D) case: when the number of bands increases the computation of the histogram becomes unfeasible due to the curse of dimensionality. However, the Gaussian and the EC model can be estimated easily for multiple dimensions. The difference between the Gaussian and the EC model relies in the kurtosis of the distribution, while in the Gaussian case is fixed in the EC case can be controlled with the ν parameter. By comparing the marginal distributions in the central and the right panels one can easily spot the differences between the Gaussian and the EC model. For the horizontal axes the data follows quite well the Gaussian model, i.e. the red solid line and the dashed red line are very similar in the central panel. However the Gaussian model fails when reproducing the probability for the vertical axes (central panel blue lines). Although it is not a perfect model, the EC distribution is better suited than the Gaussian distribution for describing the real distribution of the data. For instance in the case of the P_y (equivalent to $\mathbb{P}_Y$) distribution (blue lines, vertical axes) the EC description (solid blue line in third panel) is more similar to the original one (dashed blue lines) than the description given by the Gaussian distribution (solid blue line in second panel).

3.4 Kernel ACD algorithms

Previous methods are linear and depend on estimating covariance matrices with the available data, and use them as a metric for testing anomalousness. These methods are fast to apply, delineate point-wise nonlinear decision boundaries, but still rely on second-order statistics. This restricts the class of functions that can be implemented and thus the generalization capabilities of the algorithm. For instance in Fig. 3.1 the assumed joint distributions (dark green) for both Gaussian and EC models clearly differ from the *r*eal distribution (light green). Here, this issue is addressed through the theory of reproducing kernel func-

tions (Shawe-Taylor & Cristianini., 2004), which allows us to capture higher-order feature relations while still relying on linear algebra. Kernel methods are particularly robust to reduced sample sizes and high-dimensional feature spaces, situations often encountered in hyperspectral image detection problems.

Kernel methods constitute a well-known approach in machine learning. They have been mainly used for classification and regression, and not that much in anomaly and target detection. The problem has been approached mainly with discriminative and subspace methods: the support vector domain description (SVDD) –also known as one-class SVM–, the kernel OSP, and the kernel RX methods (Rojo-Álvarez et al., 2017). In this approach, previous anomaly change detection methods will be kernelized fallowing the same way as for deriving the kernel RX in (Kwon & Nasrabadi, 2005), yet the context is extended by assuming elliptically contoured distributions and parameterizations (see Table 3.1 and Eq. 3.3). Let us first start by introducing the kernelization of the RX algorithm. This method will be based on the theory of reproducing kernels following the same notation in the previous chapter. Note that in order to estimate the anomaly $\xi(\phi(\mathbf{x}_i))$, the same procedure will be followed as in the linear case but first mapping the points to the Hilbert space

$$\xi^{\mathcal{H}}(\mathbf{x}_i) = \phi(\mathbf{x}_i)(\mathbf{\Phi}^\top\mathbf{\Phi})^{-1}\phi(\mathbf{x}_i)^\top. \tag{3.4}$$

Note that one do not have access to either the samples or the covariance in the Hilbert. However note that $(\mathbf{\Phi}^\top\mathbf{\Phi})^{-1} = \mathbf{\Phi}^\top\ (\mathbf{\Phi}\mathbf{\Phi}^\top\mathbf{\Phi}\mathbf{\Phi}^\top)^{-1}\mathbf{\Phi}$. This can be easily shown by right multiplying by the term $\mathbf{\Phi}^\top\mathbf{\Phi}\mathbf{\Phi}^\top$ and applying some linear algebra. By substituting in eq. (3.4) one get

$$\xi^{\mathcal{H}}(\mathbf{x}_i) = \phi(\mathbf{x}_i)\mathbf{\Phi}^\top\ (\mathbf{\Phi}\mathbf{\Phi}^\top\mathbf{\Phi}\mathbf{\Phi}^\top)^{-1}\mathbf{\Phi}\phi(\mathbf{x}_i)^\top.$$

In this equation one can replace all dot products by reproducing kernel functions using the represent theorem (Shawe-Taylor & Cristianini., 2004), and hence

$$\xi^{\mathcal{H}}(\mathbf{x}_i) = \xi(\phi(\mathbf{x}_i)) = \mathbf{k}_i(\mathbf{K}\mathbf{K})^{-1}\mathbf{k}_i^\top, \tag{3.5}$$

where $\mathbf{k}_i = [K(\mathbf{x}_i,\mathbf{x}_1),\ldots,K(\mathbf{x}_i,\mathbf{x}_n)] \in \mathbb{R}^{1\times n}$ contains the similarities between $\mathbf{x}_i$ and all training data, $\mathbf{X}$, and $\mathbf{K} \in \mathbb{R}^{n\times n}$ stands for the kernel matrix containing all training data similarities. Note that, as in the linear RX method, the KRX also requires centering the

data (now in $\mathcal{H}$), which can be easily done[1]. Hereafter is assumed that all kernel matrices are centered. The solution may need extra regularization $\xi^{\mathcal{H}}(\mathbf{x}_i) = \mathbf{k}_i(\mathbf{KK} + \lambda \mathbf{I}_n)^{-1}\mathbf{k}_i^\top$, $\lambda \in \mathbb{R}^+$. Therefore the kernel version of Eq. (3.2) is:

$$\mathcal{A}_{\mathcal{G}}^{\mathcal{H}}(\mathbf{x}_i, \mathbf{y}_i) = \xi^{\mathcal{H}}(\mathbf{z}_i) - \beta_x \xi^{\mathcal{H}}(\mathbf{x}_i) - \beta_y \xi^{\mathcal{H}}(\mathbf{y}_i).$$

By following a similar procedure for Eq. (3.3), one obtains kernel versions of the elliptically-contoured linear solution:

$$\begin{aligned} \mathcal{A}_{\mathrm{EC}}^{\mathcal{H}}(\mathbf{x}_i, \mathbf{y}_i) &= (2d+\nu)\log\left(1 + \frac{\xi^{\mathcal{H}}(\mathbf{z}_i)}{\nu}\right) \\ &- \beta_x(d+\nu)\log\left(1 + \frac{\xi^{\mathcal{H}}(\mathbf{x}_i)}{\nu}\right) \\ &- \beta_y(d+\nu)\log\left(1 + \frac{\xi^{\mathcal{H}}(\mathbf{y}_i)}{\nu}\right), \end{aligned}$$

Note that in the case of $\beta_x = \beta_y = 0$, the algorithm reduces to kernel RX which was previously introduced in (Kwon & Nasrabadi, 2005).

Figure 3.2 shows the results of different ACD methods for the illustrative example presented in Fig. 3.1. Different thresholds over the anomaloussnes function, $\mathcal{A}$, are represented as contour lines. Each method obtains different decision boundaries. The ideal situation would be to have a surface where the green points are surrounded by a contour line and the yellow points are outside of the contour line. Note that this is a complex problem where no perfect solution can be achieved since the anomalous (yellow points) and non anomalous (green points) pixels are overlapped. Here, and through this context, the results will be summarized using the value of the area under the curve (AUC) of the detection receiver operating characteristic (ROC) curves. Bigger AUC means better detection of the anomalous change. As an illustration, a close look can be taken to the results for the method that achieves higher AUC, the K-EC-YX. The shape of the surface tries to keep inside the green points (although some orange points are also included). In general one can see that the kernel methods obtain better results than their linear counterpart. Note that the flexibility of the solutions is different for the Gaussian, EC, and the kernel based methods. The surfaces are direct consequence of the probabilistic model assumed, for instance in the case of RX for Gaussian and EC assumptions the surfaces are equivalent to the probabilistic distributions of $\mathbb{P}_{X,Y}$ in Fig. 3.1. It is clear that the kernel versions have much more capacity to non-linearly adapt the decision surface to the problem.

[1] Centering in feature space can be easily done implicitly via the simple kernel matrix operation $\tilde{\mathbf{K}} \leftarrow \mathbf{HKH}$, where $H_{ij} = \delta_{ij} - \frac{1}{n}$, δ represents the Kronecker delta $\delta_{i,j} = 1$ if $i = j$ and zero otherwise.

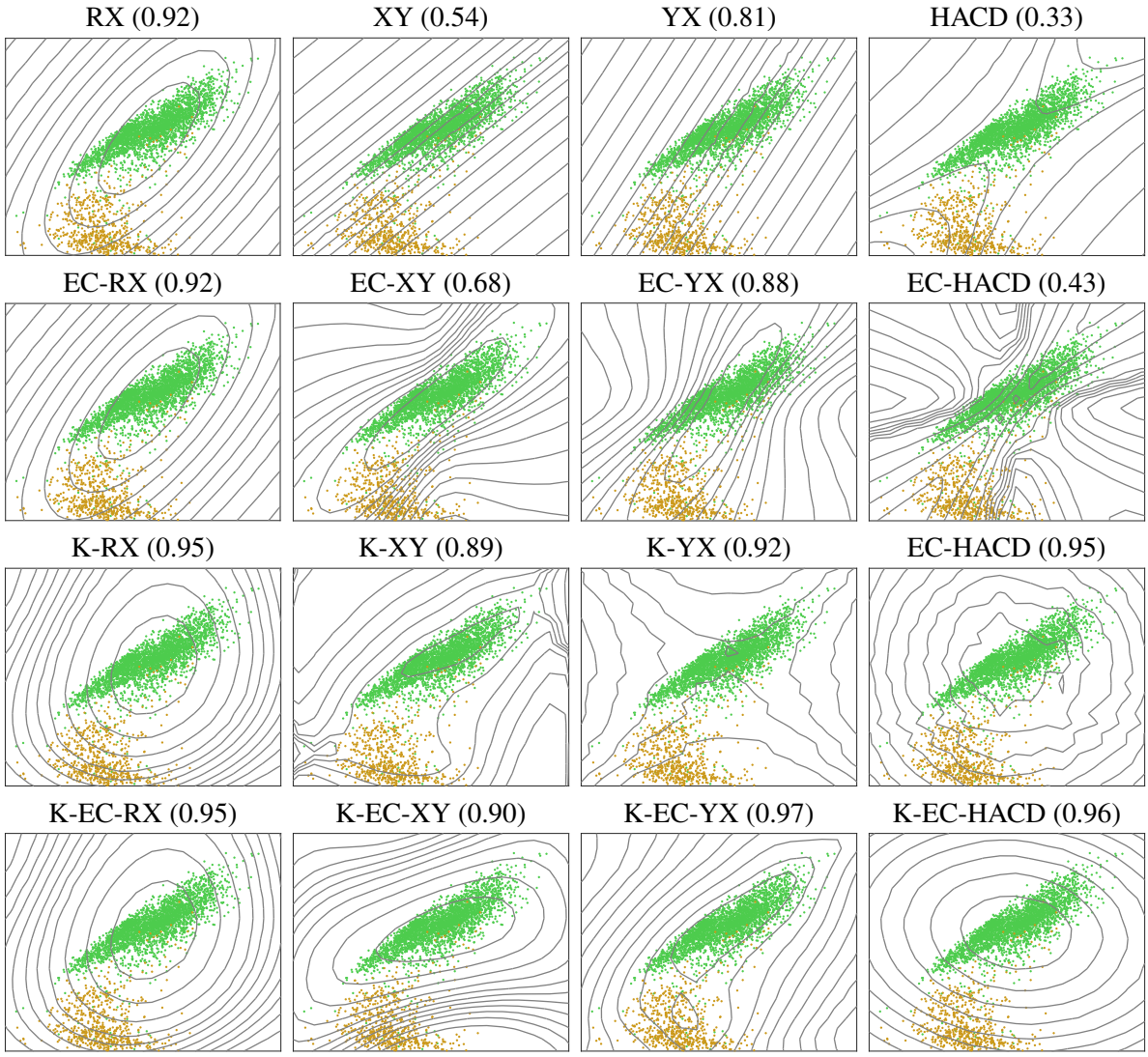


Figure 3.2: Illustration of the anomalous detection surfaces for each method. The toy example represent exclusively the band 9 of Sentinel-2 sensor. The amount of anomalies (i.e. bigger $\mathcal{A}$) is indicate by level curves. Green dots represent the non-anomalous data, while the yellow points are the anomalous data. Overall area under curve (AUC) of the receiver operating characteristic (ROC) values are given in parenthesis.

3.5 Experimental Results

This section analyzes the proposed methods. In order to test the robustness of the results, tests were performed in several simulated and real examples of pervasive and anomalous changes. The performance of the methods were evaluated by using the AUC of ROC curves.

Three experiments were performed in different datasets with complexity and control on the analyzed changes. First, an experiment was performed where the kind the anomalous

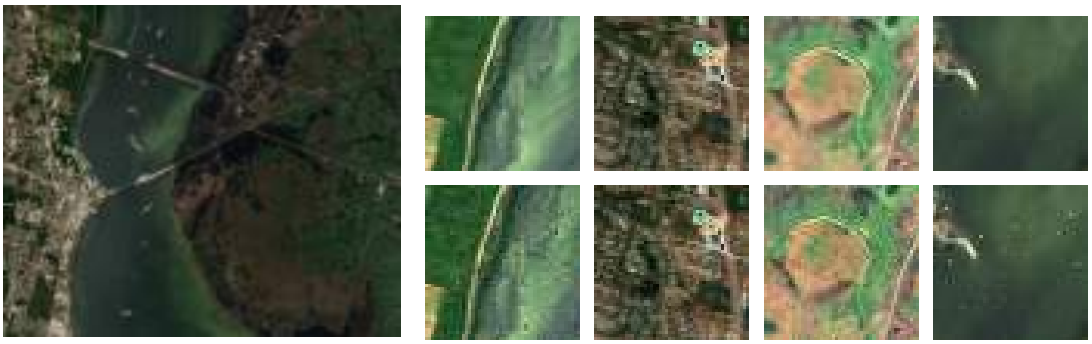


Figure 3.3: Color composite of the hyperspectral image from AVIRIS sensor (left panel), and the simulated changes (right grid panel). The original (leftmost) image is used to simulate an anomalous change image (rightmost) by adding Gaussian noise and randomly scrambling 1% of the pixels.

change was controlled in a synthetic scenario. The second experiment deals with data where the changes were real but controlled, since they were manually introduced in the scene using black tarps. Finally, the third battery of experiments deal with natural changes related to floods, droughts and man-made changes.

The Matlab implementations of the all methods have been performed. Moreover, a available database has been made with the labeled images employed in the third experiment publicly available here: `http://isp.uv.es/kacd.html`.

3.5.1 Experiment 1: Simulated Changes

This experiment is devoted to analyzing the capacity of the methods to detect pervasive and anomalous changes in simulated data by reproducing the simulation framework used in (Theiler, 2008). The data set (see Fig. 3.3) is an AVIRIS 224-channel image acquired over the Kennedy Space Center (KSC), Florida, on March 23rd, 1996. The data was acquired from an altitude of 20 km and has a spatial resolution of 18 m. After removing low SNR and water absorption bands, a total of 176 bands remain for analysis. More information can be found at http://www.csr.utexas.edu/.

Here no further dimensionality reduction was performed with PCA and, instead, work directly with the SNR filtered hyperspectral data. *Pervasive changes* are simulated by adding Gaussian noise with 0 mean and 0.1 standard deviation to all the bands and all the pixels. The image with the added noise is taken as the second image. *Anomalous changes* are produced by scrambling some pixels in the second image. Note that since only switching the position of pixels the global distribution of the image does not change. Since the methods are applied pixel-wise, this yields anomalous changes that can not be detected

as anomalies in the individual images.

In this experiment, it is limited to the use of hyperbolic detectors (HACD), i.e. $\beta_x = \beta_y = 1$, that have shown improved performance for this particular experiment (Theiler et al., 2010). All the involved parameters (estimated covariance $\mathbf{C}_z$ and kernel $\mathbf{K}_z$, ν for the EC methods, lengthscale σ parameter for the kernel versions) were tuned through standard cross-validation in the training set and show results on the independent test set.

This experiment uses the spectral angle mapper (SAM) kernel, $K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\text{acos}(\mathbf{x}_i^\top \mathbf{x}_j / (\|\mathbf{x}_i\| \|\mathbf{x}_j\|))^2/(2\sigma^2))$, since it has been proven a good choice for hyperspectral images (Camps-Valls, 2016). Two parameters need to be tuned in our kernel versions: the regularization parameter λ and the kernel parameter. In this case was used $\lambda = 10^{-5}/n$ where n is the number of training samples, and used a isotropic kernel function, whose lengthscale σ parameter is tuned in the range of 0.05-0.95 percentiles of the distances between all training samples. One should note that, when a linear kernel is used, $K(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^\top \mathbf{x}_j$, the proposed algorithms reduce to the linear counterparts proposed in (Theiler et al., 2010). The SAM kernel approximates the linear kernel for high σ values, therefore results should be improved with regard the linear versions. Working in the dual (or Q-mode) with the linear kernel instead of the original linear versions can be advantageous *only* in the case of higher dimensionality than available samples, $d \geq n$.

Figure 3.4 shows the obtained ROC curves and AUC values for the linear and kernel HACD methods. The dataset was split into small training sets of only 100 and 500 pixels, and results are given for 3000 test samples. The main conclusions are that 1) the kernel versions improve upon their linear counterparts (between 13-26% in Gaussian and 1-5% in EC detectors); 2) the EC variants outperform their Gaussian counterparts, especially in the low-sized training sets (+30% over HACD and +18% over EC-HACD in AUC terms); and 3) results improve for all methods when using 500 training samples. The EC-HACD is very competitive compared to the kernel versions in terms of AUC, but still the proposed K-EC-HACD leads to longer tails of false positive detection rates (right figure, inset plot in log-scale).

3.5.2 Experiment 2: Real and enforced Changes

This experiment is designed to analyze the performance of the proposed methods on distortions that are present in real world imagery. While the distortions that are present in any given pair of image sets are location and sensor dependant, some of the more prevalent distortions are due to seasonality, look-angle, and spatial resolution. These experiments employ a very-high spatial resolution sensor that was used to image the same

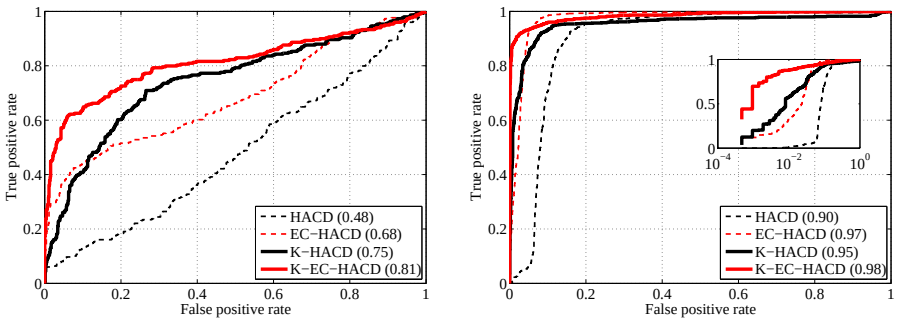


Figure 3.4: ROC curves compare the accuracy of the linear and nonlinear HACD detector based in AUC for simulated changes. On the left: the figure represent the results for 100 training samples. On the right: the figure represent the results for 500 training samples, a version in logarithmic scale is shown in the detailed plot.

target with highly varying view angles (thus, varying distortion and layover) as well as large differences in seasonality. The ability to detect anomalous changes in these highly distorted image sets illustrates the unique advantage of these types of algorithms and, in particular, the performance advantages of the proposed methods.

The experiments utilize three WorldView-2 images collected in May, August, and November of 2013. All three images (Fig. 3.5) were collected over a mixed suburban and rural area with urban residential features, roadways, rivers, and agricultural fields. The first image (May) was acquired at a relatively small off-nadir (14.0^o) angle early in the summer season. The second (Aug) and third (Nov) images were collected at much higher off nadir angles, 43.6^o and 29.3^o, respectively. In each of the final two images, one dark and one white tarp (20×20 m each) were introduced as anomalous changes.

This creates two anomalous change image sets on which to test the proposed methods with varying degree of both angular and seasonality distortions: (1) May/Aug: High off-nadir difference, moderate seasonality change; and (2) May/Nov: Moderate off-nadir difference, large seasonality change. While the white and black tarps that are introduced into the change images are highly anomalous, the spectral change is not unrepresentative of real-world problems. Additionally, the ability to more accurately model changes in highly distorted images provides a unique test case for these proposed methods.

For each experiment, 50 non-anomalous pixels were randomly selected from the stacked image sets to model the data space using the proposed algorithms. 500 randomly selected (training samples held out) non-anomalous and all anomalous pixels (May/Aug:153, May/Nov:144) were select for testing. These random selections were collected for 50 independent runs. The mean ROC curves are reported in Fig. 3.6 and the statistics for

(a) May. 2013, 14.0^o off-nadir (b) Aug. 2013, 43.6^o off-nadir (c) Nov. 2013, 29.3^o off-nadir

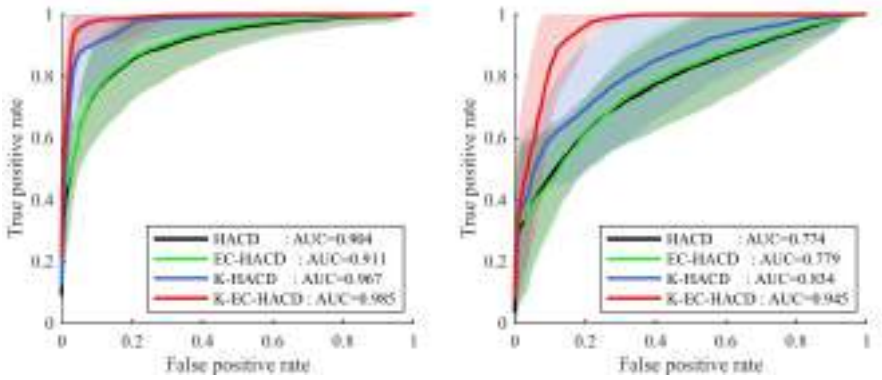
Figure 3.5: The three WorldView-2 images present a wide variety of distortions due to both seasonality and view angle. In addition to the more obvious changes in agricultural and natural vegetation, the varying view-angles result in variations in ground sample distances (GSD) of 2.0 m (May), 3.6 m (Aug), and 2.4 m (Nov).

AUC are reported in Table 3.2. As was reported earlier, the parameters ν and σ were tuned through standard cross-validation. The results are shown for independent test sets. In both of the experiments, the HACD and EC-HACD methods had almost identical average ROC curves.

Table 3.2: Area Under the Curve Statistics for the WorldView-2 View-Angle and Seasonality Experiments.

METHODS	**May-Aug Large Off-Nadir**	**May-Nov Large Seasonality**
Longmont, Colorado		
HACD	*0.90* $\pm$0.06	*0.77* $\pm$0.08
EC-HACD	*0.91* $\pm$0.06	*0.78* $\pm$0.08
K-HACD	*0.97* $\pm$0.04	*0.83* $\pm$0.11
K-EC-HACD	*0.99* $\pm$0.02	*0.95* $\pm$0.04

The parameter search for ν used in the EC-HACD method favored very large values, indicating that the data space is Gaussian and does not particularly benefit from elliptical modeling. This is most likely due to the anomalousness of the tested anomalous targets. Each of the tarp spectral signatures are highly anomalous (very dark and very bright) presenting a relatively simplified modeling space. However, the kernel methods did outperform the non-kernel methods by a statistically significant +8% and +17% as measured by mean AUC.



(a) May/Aug High Off-Nadir Experiment (b) May/Nov Large Seasonality Experiment

Figure 3.6: ROC curves for the two experiments of Section 3.5.2. The mean value of the experimental runs is plotted with the standard deviation of each detection algorithm represented by the shaded region.

3.5.3 Experiment 3: Real and Natural Changes

This experiment deals with the detection of anomalous changes that can be found naturally in a real environment.

Data collection

Pairs of multispectral images were collected, each pair consists of images taken at the same location but at different times. The images were selected in such a way that an anomalous change happened between the two acquisition times. All the images were manually labeled finding the pixels where there is an anomalous change. This step is critical and delicate since one could fall into many false alarms due to, for instance, shadows, illumination changes or natural changes in the vegetation. This is why this problem is so difficult to solve in an automatic way: for instance, one can see some areas with misclassified pixels in the prediction maps in Fig. 3.7. All images contain changes of different nature, which allows us to study how the different algorithms perform in a diversity of realistic scenarios. Table 5.1 exposes different descriptors of the images in the database. Fig. 3.7 shows the RGB composites of the pairs of images and the corresponding reference map.

Numerical comparison

Different considerations have to be taken when using the different algorithms. On the one hand the family of methods based on EC distribution involve the optimization of the ν parameter. On the other hand kernel methods involve fitting the kernel function parameters. The experiment used the classical *RBF* kernel which is well suited for multispectral images

[1] Only bands in the visible part of the spectrum were used.

Table 3.3: Images attributes in the experimentation dataset.

Images	Sensor	Size	Bands	SR
Experiment 1				
KSC	AVIRIS	512 x 614	224	18m
Experiment 2				
Longmont (May)	Worldview-2	1156 x 1563	8	2.0m
Longmont (Aug)	Worldview-2	710 x 1021	8	3.6m
Longmont (Nov)	Worldview-2	1074 x 1149	8	2.4m
Experiment 3				
Argentina	Sentinel-2	1257 x 964	12	10m-60m
Australia	Sentinel-2	1175 x 2031	12	10m-60m
California	Sentinel-2	332 x 964	12	10m-60m
Poopo Lake	MODIS[2]	326 x 201	7	250m-1km
Denver	QuickBird	500 x 684	4	1m-4m

and has only one parameter, σ. Also, the experiment have been performed using also the SAM and the polynomical kernels, however results (not shown) were worse than for the *RBF* kernel. In addition an extra parameter λ has to be fitted to regularize the matrix inversion. Selecting properly all these three parameters is an issue. An ideal situation would be having a rule of thumb to choose them. Preliminary experiments have been performed to explore the applicability of several existing rules to estimate the σ parameter. For the different images and problems faced in this section the heuristics was applied and tried to find an heuristic for the ν and λ parameters. In particular ten different heuristics were investigated: average distance between all samples, median of the distance between all samples, squared root of the dimensionality times variance per dimension averaged, median of the Silverman's rule (Silverman, 1986), median of the Scott's rule per feature (Scott, 2010), maximum likelihood density estimation, maximum Bayes estimate, maximum entropy estimate, average estimate of marginal kernel density estimate, and kernel density estimation using Gaussian kernel. While some of them have good performance for particular problems none of the rules was useful in general (results not shown). This is a usual problem in ACD where, for instance, instead of setting a particular anomaly threshold, it is usual to compute the ROC curve where all the thresholds are evaluated (Theiler et al., 2010). Instead of using a different ROC curve for each parameter the problem was simplified by adopting a cross-validation scheme to fit all the involved parameters: σ, λ, and ν. Note that not only the kernel methods but also the linear EC methods have hyper parameters to fit. A realistic scenario was adopted where one only need to have labels for a small region. One advantage to use this idea is that once the best

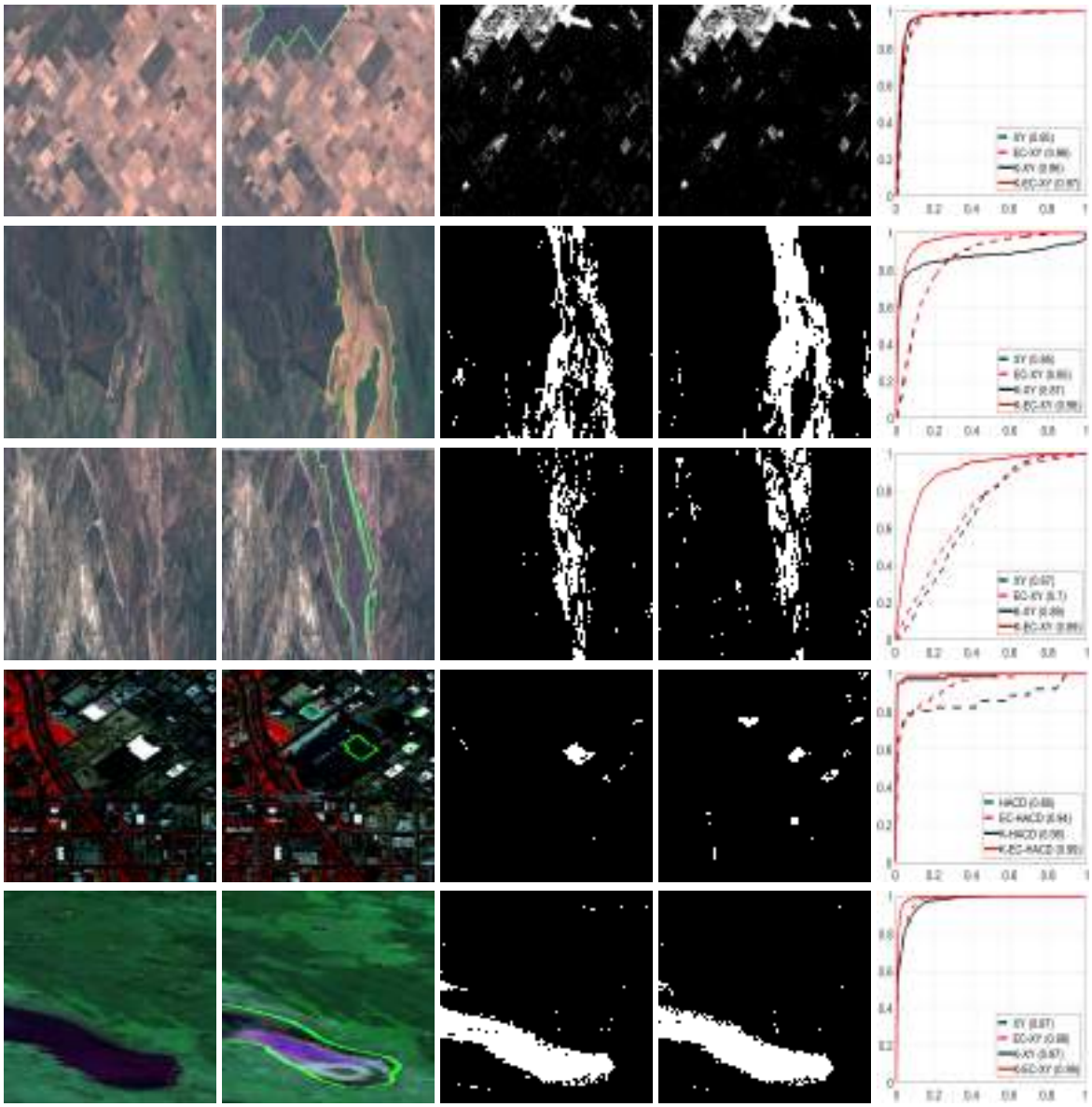


Figure 3.7: Images with *natural* anomalous changes, predictions maps and ROC curves. First row: area burned in Argentina between the months of July and August 2016, anomalous samples represent 7.5%. Second row: natural floods caused by Cyclone Debbie in Australia 2017, anomalous samples represent 17.35%. Third row: consequences of the fire in a mountainous area of California (USA), anomalous samples represent 11.33%. Fourth row: Quickbird multispectral images acquired over Denver city (USA) where appears an urbanized area, anomalous samples represent 1.6%. Last row: drying of Poopo Lake in Bolivia at the end of 2015, anomalous samples represent 11.7%. First column: images without anomalous changes. Second column: images with anomalous changes and their corresponding labels surrounded with green. Third column: prediction map using the best linear method. Fourth column: prediction map using the best kernel method. Last column: ROC curves and AUC values for the best detectors.

parameters are known in a specific region, you can apply this parameter directly without need to use cross-validation in similar scenarios. In particular, one half of the image was

used for training and obtaining the best parameters, and the other half of the image as test set. The same procedure was used for all the algorithms.

For each pair of images, they were split into two parts, and one was used for training and one for testing. The best parameters were selected by grid search in a cross-validation scheme, using 1000 training samples and 4000 validation samples randomly selected form the training set. Each method implies different set of parameters. For the ν parameter, 100 points were explored logarithmically spaced between $[10^{-5}, 10^{10}]$. For σ parameter was explored around the heuristic of the mean of the Euclidean distance between pairs of points (which was the most successful in the preliminary experiments), a grid was made by taking 60 logarithmically spaced points respectively between $[10^{-3}, 10^{3}]$ multiplied by the heuristic value. For the λ parameter, 30 values were used logarithmically spaced between $[10^{-10}, 10^{2.5}]$. Note that these methods do not give a classification but anomalousness value for each pixel. In order to provide a classification map, a particular discrimination threshold (value from which it is decided whether each pixel is an anomalous change or not) should be chosen. It is customary to provide the ROC curves. These curves represent the results of applying a binary classifier to the output of the methods for different threshold values (from more to less restrictive). Each point on the curve is the relationship between true positive and false positive corresponding to the solution provided when applying a particular threshold to the whole dataset. ROC analysis is usually employed to compare models. Here, the parameters of the different methods were optimized to maximize the AUC, in the training set (upper part of the image) and use the best parameters for the validation set (bottom part of the image).

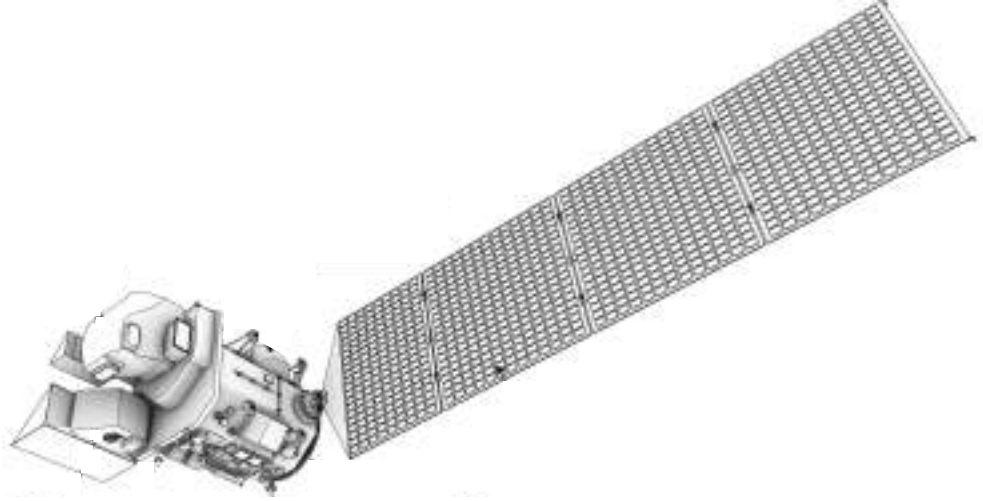
In Fig. 3.7 the ROC curves for the best method in AUC terms and Table 3.4 summarizes all AUC values for all images and methods. Fig. 3.7 compares the ability of the best linear method against the best kernel method when using the optimal threshold. Again kernel methods produce maps with less both false positives and negative alarms. As a summary, the kernel version achieves the best results in all the images when compared with its linear counterpart. Although the XY family seems to work better for the K-EC-ACD method, of the 16 detectors under study there is not an overall winner for all the families since each detector has its own characteristics (that can relatively fit data particularities), and the parameters are adjusted according to the type of image. The K-ACD version achieved better performance in both over the linear ACD, and over the linear EC-ACD. And the K-EC-ACD versions have better performance than the rest. For each type of detector (i.e. RX, XY, YX, or HACD) the AUC values can be ranked as: K-EC-ACD $\geqslant$ K-ACD $\geqslant$ EC-ACD $\geqslant$ ACD.

Table 3.4: AUC results for all five images. First and second best values for each image and each member of the family are in bold. We provide the mean and the standard deviation for ten different trials, values marked with (†) had an outlier so we give the median instead of the mean. Values marked with (•) represent the best overall result for all methods.

METHODS	RX	YX	XY	HACD
ARGENTINA				
ACD	*0.88* ±0.008	*0.86* ±0.010	**0.95** ±0.004	**0.93** ±0.007
K-ACD	**0.93** ±0.009	**0.94** ±0.007	**0.95** ±0.011	**0.93** ±0.005
EC-ACD	*0.88* ±0.008	*0.86* ±0.010	*0.95* ±0.004	*0.93* ±0.006
K-EC-ACD	**0.93** ±0.009	**0.94** ±0.008	• **0.96**±0.008	**0.95** ±0.007
AUSTRALIA				
ACD	*0.79* ±0.019	*0.79* ±0.018	**0.83** ±0.015	*0.79* ±0.012
K-ACD	**0.92** ±0.010	**0.82** ±0.019	**0.83** ±0.049	**0.89** ±0.010
EC-ACD	*0.79* ±0.019	*0.80* ±0.018	*0.83* ±0.015	*0.80* ±0.012
K-EC-ACD	**0.92** ±0.010	**0.86** ±0.016	• **0.95** ±0.008	**0.87** ±0.038
CALIFORNIA (USA)				
ACD	*0.50* ±0.015	*0.59* ±0.017	*0.65* ±0.018	*0.81* ±0.014
K-ACD	**0.61** ±0.024	**0.71** ±0.048	• **0.85** ±0.022	**0.84** ±0.013
EC-ACD	*0.50* ±0.015	*0.59* ±0.016	*0.66* ±0.024	*0.82* ±0.016
K-EC-ACD	**0.61** ±0.024	**0.71** ±0.047	• **0.85** ±0.022	**0.84** ±0.013
DENVER (USA)				
ACD	*0.95* ±0.013	**0.94** ±0.014	*0.82* ±0.059	*0.75* ±0.058
K-ACD	**0.96** ±0.023	†**0.94** ±0.050	**0.87** ±0.017	**0.96** ±0.017
EC-ACD	*0.95* ±0.013	**0.95** ±0.011	*0.88* ±0.027	*0.89* ±0.023
K-EC-ACD	**0.96** ±0.019	†**0.95** ±0.037	• **0.97** ±0.018	• **0.97** ±0.018
POOPO LAKE (BOLIVIA)				
ACD	• **0.99** ±0.002	**0.98** ±0.003	**0.96** ±0.007	*0.63* ±0.032
K-ACD	• **0.99** ±0.002	†*0.97*±0.044	**0.96** ±0.007	**0.96** ±0.005
EC-ACD	• **0.99** ±0.002	**0.98** ±0.004	*0.97* ±0.006	*0.79* ±0.034
K-EC-ACD	• **0.99** ±0.002	**0.98** ±0.013	• **0.99** ±0.002	**0.98** ±0.004

3.6 Specific contributions

This chapter presented an extension of the family of ACD methods provided in (Theiler et al., 2010) to their nonlinear counterparts based on kernel methods. The introduced methods generalize the previous ones and provide more flexible mappings to account for higher-order feature dependencies. The robustness of the proposed methods have been tested in different scenarios, including simulated, forced and realistic changes (e.g. floods, droughts and burned areas). The results of the proposed methods are better than the linear ones in all cases, demonstrating that they can be used in multiple situations. This opens up the option to use the proposed methods not only for the tested situations but also in other problems. A working implementation of all 16 methods as well as a set of labeled images have been provided, which can be used by other researchers to test ACD methods.



4. Efficient Nonlinear RX Anomaly Detectors

Contents

This chapter is partially based on the paper published in IEEE Geoscience and Remote Sensing Letters. The authors: José A Padrón Hidalgo, Adrián Pérez-Suay, Fatih Nar and Gustau Camps-Valls. Issue: 17, pages:1-5, year: 2020. Journal Impact Factor (3.83).

4.1 Summary

Current anomaly detection algorithms are typically challenged by either accuracy or efficiency. More accurate nonlinear detectors are typically slow and not scalable. In this approach, two families of techniques are proposed to improve the efficiency of the standard kernel Reed-Xiaoli (RX) method for anomaly detection by approximating the kernel function with either *data-independent* random Fourier features or *data-dependent* basis with the Nyström approach. All methods are compared for both real multi- and hyperspectral images. It is showed that the proposed efficient methods have a lower computational cost and they perform similar (or outperform) the standard kernel RX algorithm thanks to their implicit regularization effect. Last but not least, the Nyström approach has an improved power of detection.

4.2 RX Based Anomaly Detection

Among the various AD methods proposed in the literature, one of the most frequently used anomaly detectors is the Reed-Xiaoli (RX) (Reed & Yu, 1990a). In this section, the RX method is explained and its kernelized version, the KRX anomaly detector.

4.2.1 RX Anomaly Detector

It is considered an acquired image reshaped in matrix form as $\mathbf{X} \in \mathbb{R}^{n \times d}$, where n is the number of pixels and d is the total number of channels acquired by the sensor. For simplicity, let us assume that $\mathbf{X}$ is a centered data matrix. The RX detector characterizes the background in terms of the covariance matrix $\mathbf{\Sigma} = \frac{1}{d}\mathbf{X}^\top\mathbf{X}$. The detector calculates the squared Mahalanobis distance between a test pixel $\mathbf{x}_*$ and the background as follows:

$$D_{RX}(\mathbf{x}_*) = \mathbf{x}_*^\top \mathbf{\Sigma}^{-1}\mathbf{x}_*. \tag{4.1}$$

In a global AD setting, as discussed here, $\mathbf{\Sigma}^{-1}$ can be efficiently computed using all the image pixels since the dimensionality of the image is much lower than the number of pixels ($d \ll n$). Whereas, in a local AD setting, $\mathbf{\Sigma}_p^{-1}$ needs to be computed for each image pixel p using the centered pixels in a window having an origin at that pixel (Matteoli et al., 2010).

4.3 Efficient techniques for Kernel RX

Kernel methods are able to fit nonlinear problems. As it have seen in the previous chapter, kernel methods are a possible solution because they can capture higher-order (nonlinear)

feature relations, while still using linear algebra operations (Camps-Valls et al., 2009). It is proposed using feature map and low-rank approximation approaches to improve the efficiency of the KRX detector develop in 3.4 but this time focused on anomaly detection. It is studied the following approximations to the KRX method: Random Fourier features (RRX) previously studied by the authors in (Nar et al., 2018), orthogonal random features (ORX), naive low-rank approximation (LRX), and Nyström low-rank approximation (NRX).

4.3.1 Randomized Feature Map Approaches

Random Fourier Features (RFF)

An outstanding result in the recent kernel methods literature makes use of a classical definition in harmonic analysis to the approximation and scalability (Rahimi & Recht, 2007). The Bochner's theorem states that a continuous shift-invariant kernel $K(\mathbf{x},\mathbf{x}') = K(\mathbf{x}-\mathbf{x}')$ on $\mathbb{R}^d$ is positive definite (p.d.) if and only if K is the Fourier transform of a non-negative measure. If a shift-invariant kernel K is properly scaled, its Fourier transform $p(\mathbf{w})$ is a proper probability distribution. This property is used to approximate kernel functions with linear projections on a number of D random features as $K(\mathbf{x},\mathbf{x}') \approx \frac{1}{D}\sum_{i=1}^{D} \exp(-\mathbb{i}\mathbf{w}_i^\top \mathbf{x}) \exp(\mathbb{i}\mathbf{w}_i^\top \mathbf{x}')$, where $\mathbf{w}_i \in \mathbb{R}^d$ are randomly sampled from a data-independent distribution $p(\mathbf{w})$ (Rahimi & Recht, 2007). Note that it is possible to define a $2D$-dimensional *randomized* feature map $\mathbf{z} : \mathbb{R}^d \to \mathbb{R}^{2D}$, which can be *explicitly* constructed as $\mathbf{z}(\mathbf{x}) = \frac{1}{\sqrt{2D}}[\cos(\mathbf{w}_1^\top \mathbf{x}), \sin(\mathbf{w}_1^\top \mathbf{x}), \ldots, \cos(\sin \mathbf{w}_D^\top \mathbf{x}), \sin(\mathbf{w}_D^\top \mathbf{x})]^\top$ to approximate the Radial Basis Function (RBF) kernel.

Therefore, given n data points (pixels), the kernel matrix $\mathbf{K} \in \mathbb{R}^{n\times n}$ can be approximated with the explicitly mapped data, $\mathbf{Z} = [\mathbf{z}_1 \cdots \mathbf{z}_n]^\top \in \mathbb{R}^{n\times 2D}$, and will be denoted as $\hat{\mathbf{K}} \approx \mathbf{Z}\mathbf{Z}^\top$. However, this approach is not used in Equation (3.5), which would lead to a mere approximation with extra computational cost. Instead, linear RX was executed on Equation (4.1) with explicitly mapped points onto random Fourier features, which reduces to

$$D_{RRX} = \mathbf{z}_*^\top (\mathbf{Z}^\top \mathbf{Z})^{-1} \mathbf{z}_*, \tag{4.2}$$

and leads to a nonlinear randomized RX (RRX) (Nar et al., 2018) that approximates the KRX. Essentially, the original data $\mathbf{x}_i$ was mapped into a nonlinear space through the explicit mapping $\mathbf{z}(\mathbf{x}_i)$ to a $2D$-dimensional space (instead of the potentially infinite feature space with $\phi(\mathbf{x}_i)$), and then use the linear RX formula. This allows us to control the space

and time complexity explicitly through D, as one has to store matrices of $n \times 2D$ and invert matrices of size $2D \times 2D$ only (see Table 4.1). Typically, parameter D satisfies $D \ll n$ in practical applications.

Orthogonal Random Features (ORF)

RFF has become a very practical solution for the bottleneck in kernel methods when n grows. In RFF, frequencies $\mathbf{w_i}$ are sampled from a particular pdf and they act as a basis. This, however, may lead to features that are linearly dependent thus geometrically covering less space. Imposing orthogonality in the basis can be a remedy to this issue, which has led to the Orthogonal Random Features (ORF) (Yu et al., 2016). The linear transformation matrix of ORF is $\mathbf{W}_{\text{ORF}} = \frac{1}{\sigma}\mathbf{SQ}$, where $\mathbf{Q}$ is a uniformly distributed random orthogonal matrix. The set of rows of Q forms a basis in $\mathbb{R}^d$. $\mathbf{S}$ is a diagonal matrix, with diagonal entries sampled i.i.d. from the χ-distribution with d degrees of freedom. $\mathbf{S}$ makes the norms of the rows of $\mathbf{SQ}$ and $\mathbf{W}$ (with all the frequencies of RFF) identically distributed. Theoretical results show that ORF achieves lower error than RFF for the RBF kernel (Yu et al., 2016). This approach follows the above RFF philosophy, and the final anomaly score is now:

$$D_{ORX} = \mathbf{z}_*^\top (\mathbf{Z}^\top \mathbf{Z})^{-1} \mathbf{z}_*, \tag{4.3}$$

where each frequency $\mathbf{w}_i$ is a row of $\mathbf{W}_{\text{ORF}}$ and $\mathbf{Z}$ is the matrix formed by the mappings $\mathbf{z}(\mathbf{x}_i)$ of each element in the dataset, and $\mathbf{z}_*$ is the mapping of a pixel to be tested.

Nyström Approximation

The Nyström method selects a subset of samples to construct a low-rank approximation of the kernel matrix (Williams & Seeger, 2001). This method approximates the kernel function as $\mathbf{K}(\mathbf{x}_*, \mathbf{x}) \approx \mathbf{k}_{*:r}^\top \hat{\mathbf{K}}^{-1} \mathbf{k}_{\mathbf{x}:r}$, where $\mathbf{k}_{\mathbf{x}:r}$ contains the similarities between $\mathbf{x}$ and all r points, and $\hat{\mathbf{K}} \in \mathbb{R}^{r \times r}$ stands for the kernel matrix between the points in $\hat{\mathbf{X}}$. Therefore, $\mathbf{k}_*$ can be expressed as:

$$\mathbf{k}_* \approx \mathbf{R}^\top \hat{\mathbf{K}}^{-1} \mathbf{k}_{*:r}, \tag{4.4}$$

where $\mathbf{R} \in \mathbb{R}^{r \times n}$ is a matrix which contains similarities between the points in $\hat{\mathbf{X}}$ and the points in $\mathbf{X}$. The similarities were computed using the standard RBF kernel function $K(\mathbf{x}, \mathbf{y}) = \exp(-\|\mathbf{x} - \mathbf{y}\|^2 / (2\sigma^2))$.

Using the above definition given in (4.4), the Nyström method approximates the kernel

matrix $\mathbf{K}$:

$$\mathbf{K} \approx \mathbf{R}^\top \hat{\mathbf{K}}^{-1} \mathbf{R}. \tag{4.5}$$

by plugging (4.4) and (4.5) into (3.5), one can define the low-rank approximation of KRX:

$$D_{NRX}(\mathbf{x}_*) = \mathbf{k}_{*:r}^\top \hat{\mathbf{K}}^{-1} \mathbf{R} (\mathbf{R}^\top \mathbf{M} \mathbf{R})^{-1} \mathbf{R}^\top \hat{\mathbf{K}}^{-1} \mathbf{k}_{*:r}, \tag{4.6}$$

where $\mathbf{M} = \hat{\mathbf{K}}^{-1} \mathbf{R} \mathbf{R}^\top \hat{\mathbf{K}}^{-1}$ while $\mathbf{M} \in \mathbb{R}^{r \times r}$. Since $\mathbf{R}$ is not a squared matrix $(r < n)$, it is rank deficient, and it is proposed to use the pseudoinverse instead of the inverse of $\mathbf{R}^\top \mathbf{M} \mathbf{R}$. By doing this, most of the terms cancel, leading to a more compact equation for the NRX:

$$D_{NRX}(\mathbf{x}_*) = \mathbf{k}_{*:r}^\top (\mathbf{R}\mathbf{R}^\top)^\dagger \mathbf{k}_{*:\mathbf{r}}. \tag{4.7}$$

Note that NRX involves the inversion of an $r \times r$ matrix which is much more efficient compared to KRX. In addition, the Nyström approach is more generic than using random Fourier feature approaches, as it allows one to approximate all positive semidefinite kernels, not just shift-invariant kernels. Furthermore, this approximation is data-dependent (i.e. the basis functions are a subset of estimation data itself) which could translate into better results (Yang et al., 2012).

Connection to reduced-set methods

Reduced-set techniques were successfully used to obtain sparse kernel methods and low rank approximations of multivariate kernel methods (Arenas-Garcia et al., 2013). This methodology can be applied to approximate KRX which leads to equation (4.7). In this approach, the data matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$ is subsampled into $\hat{\mathbf{X}} \in \mathbb{R}^{r \times d}$, $r \ll n$, and mapped into $\hat{\mathbf{\Phi}} \in \mathbb{R}^{r \times d_{\mathcal{H}}}$, which, by using (3.4), it is obtained the LRX formula:

$$D_{LRX}(\mathbf{x}_*) = \phi(\mathbf{x}_*)^\top \hat{\mathbf{\Phi}}^\top (\hat{\mathbf{\Phi}} \mathbf{\Phi}^\top \mathbf{\Phi} \hat{\mathbf{\Phi}}^\top)^{-1} \hat{\mathbf{\Phi}} \phi(\mathbf{x}_*). \tag{4.8}$$

Identifying $\mathbf{k}_{*:r} = \hat{\mathbf{\Phi}} \phi(\mathbf{x}_*)$ and $\mathbf{R} = \hat{\mathbf{\Phi}} \mathbf{\Phi}^\top$, (4.8) leads to:

$$D_{LRX}(\mathbf{x}_*) = \mathbf{k}_{*:r}^\top (\mathbf{R}\mathbf{R}^\top)^{-\mathbf{1}} \mathbf{k}_{*:\mathbf{r}}, \tag{4.9}$$

which just differs from (4.7) in the inverse of $\mathbf{R}\mathbf{R}^\top$, and when $\mathbf{R}$ is full rank they are the same. In the following and in the experiments, will be used only D_{NRX} instead of D_{LRX} as both are mathematically equivalent.

4.3.2 Space and Time Complexity

Table 4.1 gives the theoretical computational complexity of the benchmark methods (RX, KRX, SRX) and proposed methods (RRX, ORX, NRX) presented in this Thesis. In this approach, $d < D < r \ll n$ is assumed since it is aimed to deal with big data settings. Besides, KRX becomes sufficiently efficient when n is small, e.g. $n < 4000$ for a 200×200 image. As seen in Table 4.1, RX provides the best efficiency; thus, it should be employed for scenes where the data is Gaussian distributed. However, KRX and the proposed KRX approximations should be used for nonlinear distributions. Clearly, KRX is the least efficient compared to the proposed approximations, and it is also not applicable to big data. Feature map methods, e.g. RRX and ORX, provide the best computational efficiency for nonlinear (i.e non-Gaussian) distributions, while low-rank approximation methods, e.g. LRX and NRX, are also efficient yet relatively slower compared to the feature map methods. Thus, one should choose the proper method based on the image distribution characteristics (D. Manolakis & Rossacci, 2007; Keshava, 2004), detection performance requirements, and computational resource limitations. These conclusions are assessed experimentally in the following section.

Table 4.1: Memory and time complexity for all methods. T is transformation of image into a nonlinear space. C is matrix (covariance, kernel etc.) and C^{-1} is its inverse.

	Space		Time			
Method	T	C^{-1}	T	C	C^{-1}	AD
RX	–	d^2	–	nd^2	d^3	nd^2
RRX & ORX	nD	D^2	ndD	nD^2	D^3	nD^2
NRX	nr	r^2	ndr	nr^2	r^3	nr^2
KRX	n^2	n^2	n^2d	n^3	n^3	n^3

4.4 Experimental Results

This section analyzes the performance of the proposed nonlinear RX anomaly detection methods. Tests have been performed in four real examples, and tested robustness using the area under curve (AUC) of receiver operating characteristic (ROC) curves. It is provided illustrative source code for all methods in http://isp.uv.es/code/fastrx.html

4.4.1 Data collection and experimental setup

Multispectral and hyperspectral images were acquired by the Quickbird and AVIRIS sensors. Fig. 4.1 showcases the scenes used in the experiments. The AD scenarios consider

(a) (b) (c) (d)

Figure 4.1: Images with anomalies (outlined in yellow) in four scenarios: (a) consequences of the hot spots corresponding to latent fires at the World Trade Center (WTC) in NYC (extension of anomalous pixels represents the 0.23% of the image), (b) urban area where anomalies are vehicles in Gainesville city (0.52%), (c) Quickbird multispectral images acquired over Denver, the anomalies are roofs in an urbanized area (1.6%), and (d) a beach scene where the anomalies are ships captured by AVIRIS sensor (2.02%) over San Diego, USA.

Table 4.2: Images attributes used in the experimentation dataset.

Images	Sensor	Size	Bands	Resolution
WTC	AVIRIS	200 x 200	224	1.7 m
Gainesville	AVIRIS	100 x 100	190	3.5 m
Denver	Quickbird	500 x 684	4	1m-4m
San Diego	AVIRIS	100 x 100	193	7.5 m

anomalies related to: latent fires, vehicles, urbanization (roofs) and ships (Guo et al., 2016; Kang et al., 2017; Padrón-Hidalgo et al., 2019). Table 5.1 summaries relevant attributes of the datasets such as sensors, spatial and spectral resolution. Parameter estimation is required for the RX, KRX, RRX, ORX and NRX. First of all, the KRX method and its proposed variants involve the optimization of the σ parameter of the RBF kernel. For the feature map approaches (RRX and ORX), the number of basis, D, parameter should be optimized. Whereas, for low-rank approximations (NRX), the number of random sub-samples, r, parameter should be optimized. A cross-validation scheme was adopted to select all the involved parameters: number of Fourier basis D, rank r, and RBF parameter σ. It is selected the parameters using different data sizes ranging between 10^3 and 3×10^4 samples.

4.4.2 Numerical comparison

The averaged AUC results were reported for all cases with 1000 runs (standard deviations were always lower than 3×10^{-3} and hence are not reported). Figure 4.2 shows that nonlinear methods improve detection over the linear RX and NRX outperforms the other approximations in three out of the four images. The AUC values of KRX are related to the inversion of a relatively big matrix. This raises the issues of poorly estimated matrices

(with a huge condition number) which are also computationally expensive to invert ($\mathcal{O}(n^3)$). However, all the proposed fast kernel RX methods have the advantage of solving both issues. Firstly, thanks to the cross-validation procedure, an estimate of the optimal number of features (RRX, ORX) or samples (NRX) can be obtained, allowing to better capture the intrinsic dimensionality of the mapped data. In a previous work (Morales-Álvarez et al., 2018), authors showed that optimizing the number of frequencies in random Fourier features approaches acts as an efficient regularizer leading to better estimates with a reduced number of frequencies needed. And secondly, fast versions are able to obtain better performance in AUC metric at a fraction of the cost (see Fig. 4.2).

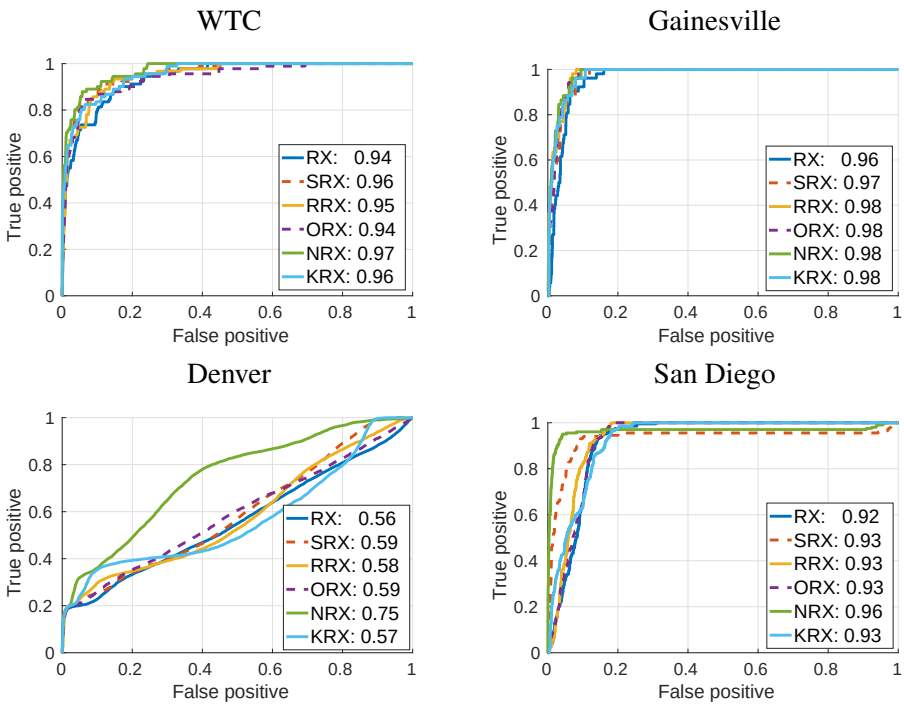


Figure 4.2: ROC curves in linear scale for all scenes. Numbers in legend display the AUC values for each method.

4.4.3 On the computational efficiency

Figure 4.3 illustrates the trade-off between the computational execution time and the AUC. The crosses indicate different values of rank (D or r parameters) in the set $\{50, 100, 200, 400, 500\}$ and the number of pixels was fixed to $n = 3000$. The optimal parameters estimated for KRX are used for the fast approaches. KRX has the best AUC values in all the images. NRX and SRX are more sensitive to rank values. RRX and ORX are almost insensitive to the rank but results do not improve when the rank increases, thus limiting their performance. The combination of lower spectral information and the ambiguity of the class (note

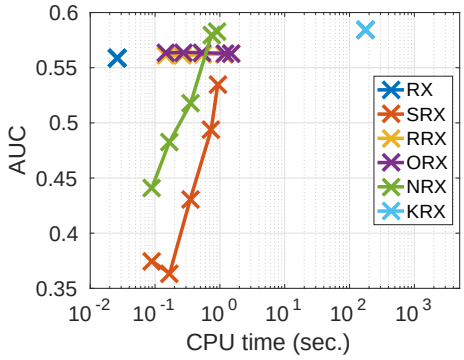


Figure 4.3: CPU execution time versus the AUC values for $n = 3000$ pixels, crosses corresponds to different rank values for Denver image.

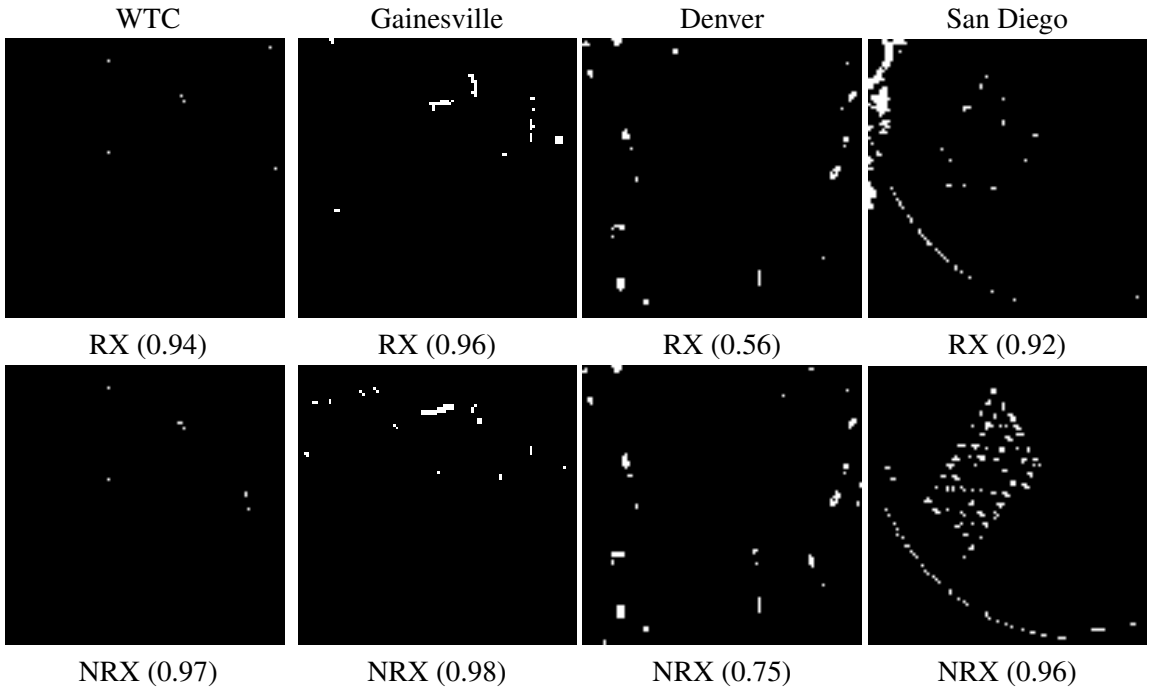
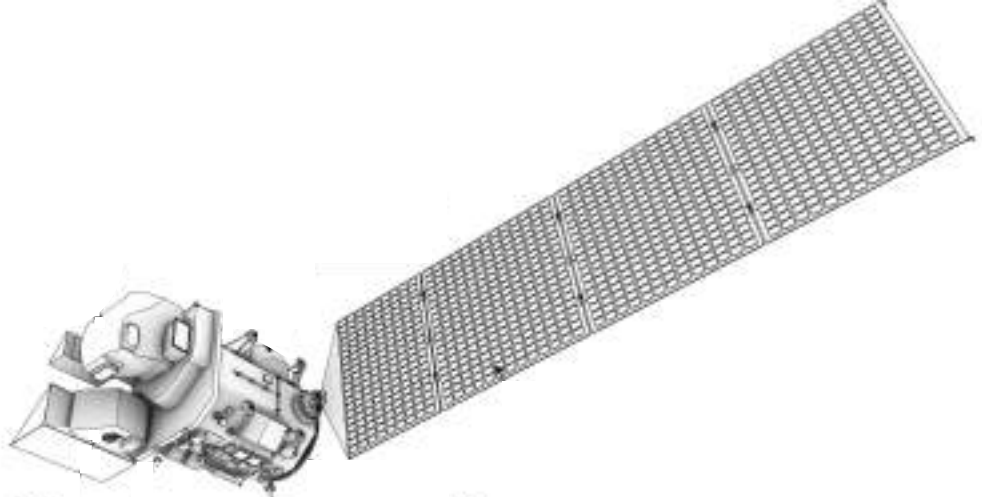


Figure 4.4: Anomaly detection maps for best thresholds (top: the best linear RX (AUC) results, bottom: best nonlinear RX (AUC) method).

that the anomaly class ‘urbanized’ can be confused with a pervasive class ‘urban’) makes the Quickbird scene a very difficult problem (lower AUCs). In this situation, as the rank parameter r for the NRX method grows, it approximates the KRX algorithm. In Figure 4.4, the RX detector (top row) is shown against the best detector obtained (bottom row). The best result in AUC was achieved by the NRX in all the images. It is worth mentioning the good results in detection achieved by the NRX in all the scenes, which can be visually compared the linear version.

4.5 Specific contributions

In this chapter, the goal was on improving the space (memory) and time (cost) of the KRX anomaly detector. Kernel-based anomaly detectors provide excellent detection performance since they are able to characterize non-linear backgrounds. In order to undertake this challenge, has been proposed to use efficient techniques based on random Fourier features and low-rank approximations to obtain improved performance of the KRX algorithm. Among all methods, the Nyström and the equivalent low-rank (LRX) approximation achieves the best results and yields a more efficient and accurate non-linear RX method to be applied in practice.



5. MULTIVARIATE GAUSSIANIZATION

Contents

This chapter is partially based on the paper ' Unsupervised Anomaly and Change Detection with Multivariate Gaussianization' submitted to IEEE Transactions on Geoscience and Remote Sensing. The authors: José A Padrón Hidalgo, Valero Laparra and Gustau Camps-Valls. Submitted. Journal Impact Factor (5.85).

5.1 Summary

Anomaly detection is a field of intense research in remote sensing image processing. Identifying low probability events in remote sensing images is a challenging problem given the high-dimensionality of the data, especially when no (or little) information about the anomaly is available a priori. While plenty of methods are available, the vast majority of them do not scale well to large datasets and require the critical choice of some (very often critical) hyperparameters. Therefore, unsupervised detection methods with an efficient use of memory become necessary, especially now with the data deluge problem. In this approach, an unsupervised method is proposed for detecting anomalies and changes in remote sensing images by means of a multivariate Gaussianization methodology that allows to estimate multivariate densities accurately, a long-standing problem in statistics and machine learning. The methodology transforms arbitrarily complex multivariate data into a multivariate Gaussian distribution. Since the transformation is differentiable, by applying the change of variables formula one can estimate the probability at any point of the original domain. The assumption is straightforward: pixels with low estimated probability are considered anomalies. Our method is flexible enough to describe any multivariate distribution, makes an efficient use of memory, and is parameter-free. The efficiency of the method is shown in experiments involving both anomaly detection and change detection in different remote sensing image sets. For anomaly detection two approaches were proposed. The first using directly the Gaussianization transform and the second using an hybrid model that combines Gaussianization and the Reed-Xiaoli (RX) method typically used in anomaly detection. Results show that our approach outperforms other linear and nonlinear methods in terms of detection power in both anomaly and change detection scenarios, showing robustness and scalability to dimensionality and sample sizes.

5.2 Multivariate Gaussianization

The rotation-based iterative Gaussianization (RBIG) is a nonparametric method for density estimation of multivariate distributions (Laparra et al., 2011). RBIG is rooted in the idea of Gaussianization, introduced in the seminal work by (Friedman, 1987) and further developed in (Chen & Gopinath, 2000; Laparra et al., 2011), which consists of seeking for a transformation G_x that converts a multivariate dataset $\mathbf{X} \in \mathbb{R}^{\ell \times d}$ in domain X to a domain where the mapped data $\mathbf{Y} \in \mathbb{R}^{\ell \times d}$ follows a multivariate normal distribution in

domain Y, i.e. $p_Y(\mathbf{y}) \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$:

$$\begin{aligned} G_x: \quad & \mathbf{x} \in \mathbb{R}^d \quad \mapsto \quad \mathbf{y} \in \mathbb{R}^d \\ & \sim p_X(\mathbf{x}) \qquad\quad p_Y(\mathbf{y}) \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d), \end{aligned} \tag{5.1}$$

where inputs and mapped data points have the same dimensionality $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, $\mathbf{0}$ is a vector of zeros (for the means) and $\mathbf{I}_d$ is the identity matrix for the covariance of dimension d. Using the change of variable formula one can estimate the probability of a point $\mathbf{x}$ in the original domain:

$$p_X(\mathbf{x}) = p_Y(\mathbf{y}) |J_{G_x}(\mathbf{y})|, \tag{5.2}$$

where $p_X(\mathbf{x})$ is the probability distribution of the original data point $\mathbf{x}$, and $|J_f(\mathbf{y})|$ is the determinant of the Jacobian of the transformation G_x in the point $\mathbf{y}$. For this formula to work, G_x has to be differentiable, i.e. the $|J_{G_x}(\mathbf{y})| > 0, \forall \mathbf{y}$. The proposed Gaussianization method in this Thesis, RBIG, obtains a transformation G_x that fulfills this property, cf. (Laparra et al., 2011). The other part of the product is easy to compute since $p_Y(\mathbf{y})$ can be estimated since p_Y is a multivariate Gaussian by construction. Therefore RBIG can be easily applied to estimate the probability of data points in the original domain, $p_X(\mathbf{x})$.

RBIG is an iterative algorithm, where in each iteration, n, two steps are applied: 1) a set of d marginal Gaussianizations to each of the variables, $\mathbf{\Psi} = [\Phi_1, \ldots, \Phi_d]$, and 2) a linear rotation, $\mathbf{R} \in \mathbb{R}^{d \times d}$:

$$\mathbf{x}[n+1] = \mathbf{R}[n] \cdot \mathbf{\Psi}[n](\mathbf{x}[n]), \quad n = 1, \ldots, N \tag{5.3}$$

where N is the number of steps (iterations) in the sequence, $n = 1, \ldots, N$. The final transformation G_x is the composition of all performed transformations through iterations. In (Laparra et al., 2011) is demonstrated that with enough iterations the method converges and the transformed data follows finally a standardized Gaussian, i.e. $p_Y(\mathbf{y}) \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$, taking $\mathbf{y} = \mathbf{x}[N]$.

An illustration of how RBIG can be adapted to describe the distribution of remote sensing data is shown in Fig. 5.1. In this example, the dataset is taken from the Sentinel-2 image Australia (see Table 5.1 for details), which has $d = 12$ bands, and use RBIG to Gaussianize its pixel's distribution. Therefore, it can be observed that the Gaussianized data follows a Gaussian distribution. Besides, the inverse of the learned Gaussianization transformation is applied to randomly generated Gaussian points obtaining synthetic new

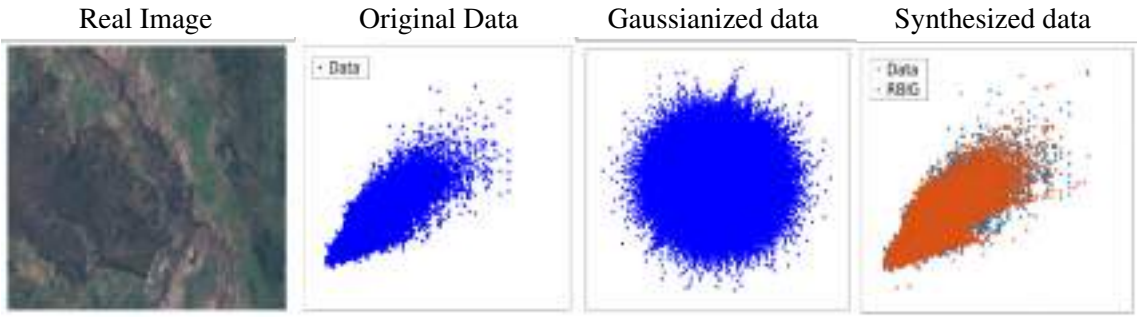


Figure 5.1: Illustration of synthesized data using RBIG approach in real images. From left to right: images in rgb composition, representation of the values for the first two bands of the image, Gaussianized data, and synthesized data.

data that follows a deemed similar distribution as the original one. This illustrates the invertibility property of RBIG, which allows us to estimate densities in the original domain and use the well-known relation between probability and anomaly to derive unsupervised density-based anomaly and change detectors.

5.2.1 RBIG for Detection of Anomalies

One of the most successful methods applied to the problem of anomaly detection is the Reed-Xiaoli (RX) method (Reed & Yu, 1990a), a successful type of matched filter. The idea behind the RX method can be interpreted in probabilistic terms (Padron-Hidalgo et al., 2021); intuitively, a data point is more anomalous when it has less probability to appear:

$$A(\mathbf{x}) \propto \frac{1}{p_X(\mathbf{x})}. \tag{5.4}$$

Actually, when the distribution is assumed to be Gaussian, $p_X \sim p_G$, this relation defines the RX method anomaly detector, i.e. $A(\mathbf{x}) \sim A_{\mathrm{RX}}(\mathbf{x})$. Actually $A_{\mathrm{RX}}(\mathbf{x})$ is equivalent to the Mahalanobis distance between the data point and the mean, i.e. $A_{\mathrm{RX}}(\mathbf{x}) = (\mathbf{x} - \mu)^\top \Sigma^{-1}(\mathbf{x} - \mu)$, where $p(\mathbf{x}) \sim \mathcal{N}(\mathbf{0}, \Sigma)$.

While RX has been widely used, it has the limitations inherent to the Gaussian distribution assumption. The use of kernel methods has been proposed to generalize the RX method to the nonlinear and non-Gaussian case (Heesung Kwon & Nasrabadi, 2004; Padron-Hidalgo et al., 2021). Kernel methods define the covariance in a higher dimensional Hilbert feature space, which in the RX method translates into replacing the covariance matrix by a kernel matrix that estimates the similarity between samples (Camps-Valls et al., 2009; Rojo-Álvarez et al., 2017). In practice this implies that correlation is substituted

by a non-linear similarity measure. Therefore the anomaly detected using the kernel RX (KRX) method can be formulated as:

$$A_{\text{KRX}}(\mathbf{x}) \propto \frac{1}{p_K(\mathbf{x})}, \tag{5.5}$$

where $p_K(\mathbf{x})$ is the distribution induced by using the kernel function instead of the covariance. The kernel RX (KRX) is an elegant extension of the RX, yet it has the problem of fitting kernel parameters and the high computational cost (as one has to invert a kernel matrix, which has cubic cost with the number of points ℓ). Whereas some heuristics exist in the literature to fit the kernel parameters, in practice one only achieves the full potential of the KRX approach by fitting the parameters after cross-validation (Padron-Hidalgo et al., 2021). This requires having access to labeled data as anomalous versus non-anomalous classes, which is not a very realistic and not even practical setting. In addition, the more useful and practical of unsupervised anomaly detection (i.e. no labeled data available) problems will be addressed. Therefore, in these comparisons, the kernel method parameter will be fitted using the most successful (and sensible) heuristic to set the Gaussian kernel lengthscale σ as the average of all distances among $\mathbf{X}$.

As an alternative to linear measures of anomalousness like in RX, or nonlinear yet implicit feature transformations with parameters to tune like in KRX, here is proposed a straightforward approach to estimate the probability density function with RBIG (sec. 5.2). This will give us a nonparametric parameter-free and efficient estimation of the data distribution. RBIG has optimal way of fitting the parameters of the distribution that do not require labeled data, and scales linearly with the data. By using RBIG to compute p_X, the method proposed is described:

$$A_{\text{RBIG}}(\mathbf{x}) \propto \frac{1}{p_{\text{RBIG}}(\mathbf{x})}. \tag{5.6}$$

An important aspect to take into account is the intrinsic characteristics of the data used to estimate the density, which has implications in the quality of the estimation. When the distribution contains even a moderate number of anomalies, an accurate density estimate will cast anomalies as regular points, i.e. non-anomalous. This vastly depends on the flexibility of the class of models used. When the model is rigid like in the RX case, this is not a problem since it cannot be adapted to the anomalies. For the KRX one can control this effect by tuning the kernel lengthscale and the regularization term, but as explained before requires labeled data. This is an important aspect to take into account mostly in

the anomaly detection scenario, where all data (included the anomalous samples) are used to estimate the density. Therefore, an hybrid model that combines the (too rigid) RX model with the (too flexible) RBIG model is proposed. The hybrid model first selects the data more likely not to be anomalous using RX and then uses this data to learn the Gaussianization transform with the RBIG model. This tries to avoid using anomalous data to train RBIG, which after all is intended to learn the background or pervasive data distribution. The number of data points selected as non-anomalous in the first step will define the trade-off between flexibility and rigidity.

5.2.2 RBIG for Change Detection

Change detection can be approached by setting thresholds on the change image (i.e. the difference between the two subsequent images for optical imagery or ratios in radar imagery) or from a purely density estimation standpoint. It will be approached it from the latter angle using RBIG. This is certainly a more challenging approach, but has several associated advantages: 1) only the first image (or all previous images before the changed one) is considered to estimate the regular/background density; 2) there is no need to corregister images since the method operates in the geometric space defined by the image, not in the spatial domain; and 3) unlike a discriminative approach, a generative model like RBIG will allow us to derive useful descriptors of the image statistics, as well as to be refined as more images are acquired.

The idea to exploit RBIG for change detection is using data coming from the first image $\mathbf{X}_1$ only to estimate the probability model and then evaluating the probability (or change score, C) for each point in the second image $\mathbf{X}_2$, as follows:

$$C(\mathbf{x}_2) \propto \frac{1}{p_{X_1}(\mathbf{x}_2)}. \tag{5.7}$$

As for the anomaly detection case, one can use different models to estimate p_{X_1}. The most widely used is the Gaussian model. As in the previous section, when assuming a Gaussian distribution for the input data, the RX method can be used here too, i.e. $C_{\mathrm{RX}}(\mathbf{y})$.

Likewise, kernel methods have been proposed to alleviate the strict assumption of Gaussian distribution (Padron-Hidalgo et al., 2021) . While different configurations were proposed in order to take into account only the anomalous changes, here one use the configuration designed for change detection. Following the idea in equation (5.7), the data of the first image ($\mathbf{X}_1$) is used to estimate the kernel and then the method is evaluated in

the second image:

$$C_{\mathrm{KRX}}(\mathbf{x}_2) \propto \frac{1}{p_{\mathrm{K}(X_1)}(\mathbf{x}_2)}. \tag{5.8}$$

Equivalently, one can use RBIG to estimate the probability of the first image and evaluate the probability in the second one:

$$C_{\mathrm{RBIG}}(\mathbf{x}_2) \propto \frac{1}{p_{\mathrm{RBIG}(X_1)}(\mathbf{x}_2)}. \tag{5.9}$$

It is important to note that, in this case, the data used to estimate the probability density does not contain anomalies (changes in this setting) so the hybrid model is not needed here.

5.3 Experimental Results

This section analyzes the performance of the proposed RBIG method for anomaly and change detection. In order to assess the robustness, tests were performed in both simulated and real scenes of varying dimensionality and sample size. The detection power of the methods were evaluated quantitatively through the Receiver Operating Characteristic (ROC) and Precision-Recall (PR) curves, along with the Area Under the Curve (AUC) scores. Besides, examples of detection maps of each method were provided to evaluate their quality by visual inspection.

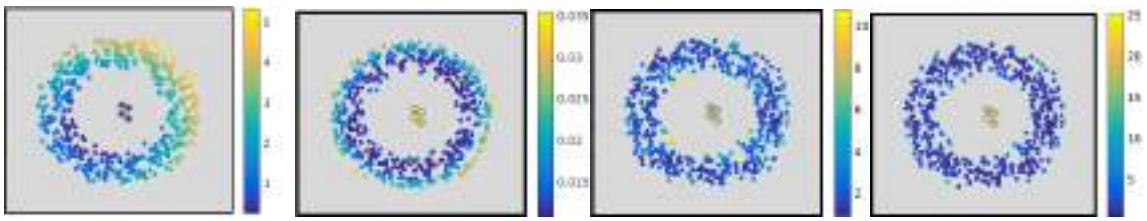
Three experiments have been performed in this setting. The first experiment is designed to illustrate the effect of the evaluated in an anomaly detection (AD) toy example. The second experiment deals with AD problem in different real scenarios: detection of air planes, latent fires, vehicles, and urbanization (roofs). The third experiment is related to evaluate the methods in change detection (CD) problems involvin floods, fires and droughts. Table 5.1 summarizes the different data sets used in the experiments. In order to ease the reproducibility, MATLAB code implementations of the all methods are provided. Moreover, a database with the labeled images used in the second and third experiments is available in https://isp.uv.es/RBIG4AD.html.

5.3.1 Experiment 1: Simulated Anomalies

The aim of this experiment is to illustrate the behavior of the proposed methods in challenging distributions exhibiting highly nonlinear feature relations. A two-dimensional dataset

Table 5.1: Images attributes in the experimentation dataset. **AD** : Anomaly Detection dataset. **CD** : Change Detection dataset.

Images	Sensor	Size	Bands	SR
AD				
Cat-Island	AVIRIS	150 x 150	188	17.2m
WTC	AVIRIS	200 x 200	224	1.7m
Texas-Coast	AVIRIS	100 x 100	204	17.2
PAVIA	ROSIS-03	150 x 150	102	1.3
CD				
Texas	Cross-Sensor	301 x 201	7	30m
Argentina	Sentinel-2	1257 x 964	12	10-60m
Chile	Landsat-8	201 x 251	12	10-60m
Australia	Sentinel-2	1175 x 2031	12	10-60m



(a) RX (b) KRX (c) RBIG (d) HYBRID

Figure 5.2: Synthetic experiment to illustrate the methods performance when detecting anomalies. The color bar shows the intensity in terms of anomaly score from dark blue (less) to yellow (more). The image (a) correspond to RX detector, image (b) is the kernel version of RX, (c) represent the RBIG method and (d) showcase the hybrid model.

was designed, where the non-anomalous data is in a circumference and the anomalous data in the middle. Figure 5.2 shows the performance of the different methods. The RX method assumption does not hold (the data is clearly non-Gaussian), hence it shows poor performance. The performance of KRX is better than RX but some false detections emerge in the outer circle, mainly related to the difficulty to select a reasonable kernel parameter. The direct application of RBIG easily identifies the anomalous points since they are far from the more dense (most probable) region. The proposed hybrid model further refines the detection since the density is estimated from pervasive data yielded by RX only.

5.3.2 Experiment 2: Anomaly Detection in Real Scenarios

Tests were performed in four real examples. Table 5.1 summarizes relevant attributes of the datasets such as sensors, spatial and spectral resolution.

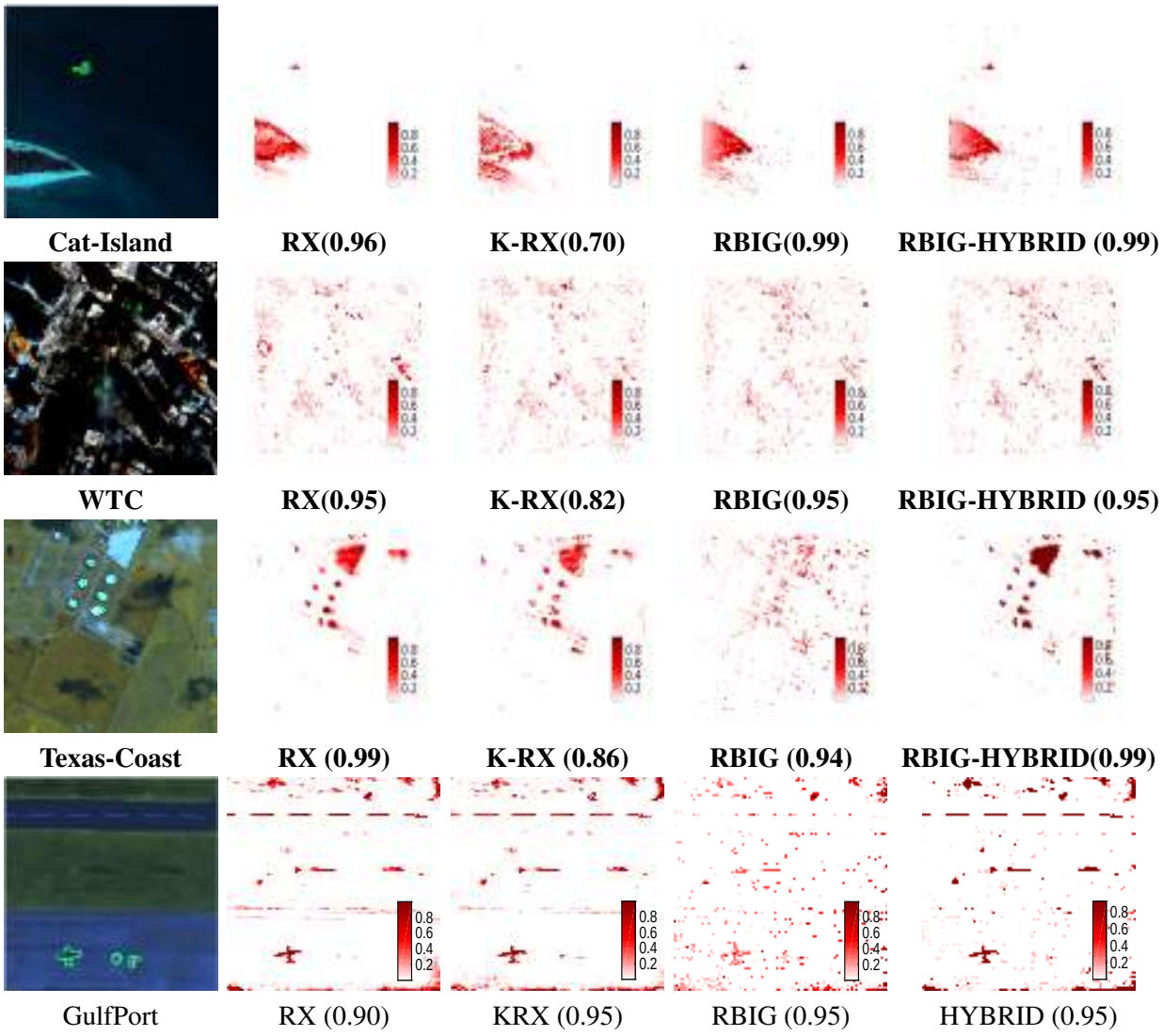
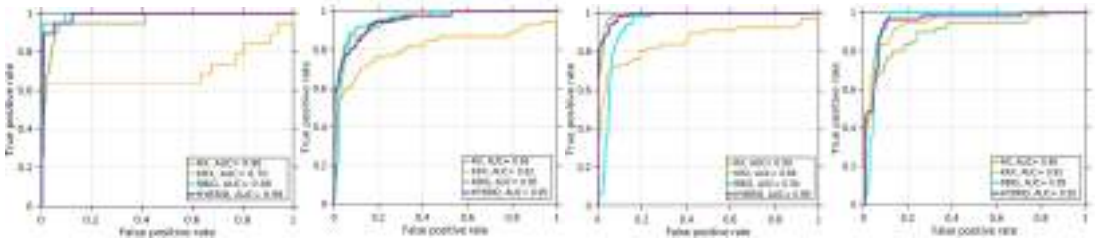


Figure 5.3: Anomaly detection predictions in four images (one per row). First column: Cat-Island, World Trade Center (WTC), Texas Coast and Pavia original datasets with anomalies outlined in green. From second column to the last column: activation maps and the AUC values (in parenthesis) for the RX, KRX, RBIG and the HYBRID models, respectively.

Data collection

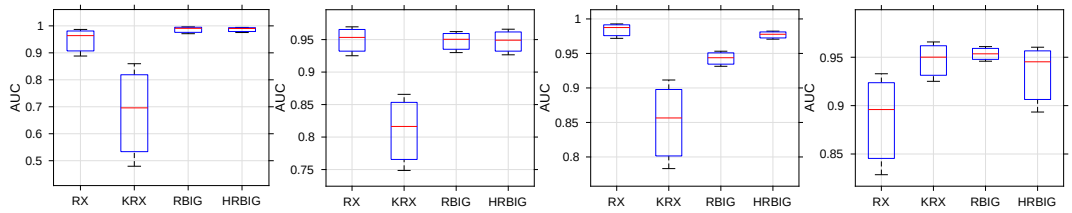
Multispectral and hyperspectral images acquired by the AVIRIS and ROSIS-03 sensors were collected. Figure 5.3 showcases the scenes used in the experiments. The AD scenarios consider anomalies related to a diversity of problems: airplane, latent fires, urbanization and vehicle detection (Guo et al., 2016; Kang et al., 2017; Padrón-Hidalgo et al., 2019).

The Cat-Island dataset corresponds to the airplane captured flying over the beach and it is considered a strange object when compared to the rest of the image (a white spot in the middle of a beach) and the percentage of anomalies represent the 0.09% of the scene.



(a) Cat-Island (b) WTC (c) Texas Coast (d) GulfPort

Figure 5.4: Anomaly detection ROC curves in linear scale for all scenes. Numbers in legend display the AUC values for each method.



(a) Cat-Island (b) WTC (c) Texas Coast (d) GulfPort

Figure 5.5: Anomaly detection results of the bootstrap experiment for 1000 experiments. AUC values and standard deviation for each method are shown as boxplot, red line represent the median value, the blue box contains 95% of the values, black lines represent the maximum and minimum values.

The World Trade Center (WTC) image was collected by the Airborne Visible Infra-Red Imaging Spectrometer (AVIRIS) over the WTC area in New York on 16 September 2001 (after the collapse of the towers in NY). The data set covered the hot spots corresponding to latent fires at the WTC, which can be considered as anomalies and it represent the 0.23% of the scene. In the Texas Coast dataset, the anomalies represent the 0.67% of the scene and the image contains roofs built on a wooded site and bright spots that reflect light which can be considered an anomaly. The GulfPort dataset correspond to a battery of airplanes taxied on the runway and the pecentage of anomalies represent the 0.60% of the scene.

Numerical and Visual Comparison

It is important to take into consideration that KRX requires the selection of some hyper-parameters, being the kernel parameter the most critical one. In order to perform a fair comparison while staying in an unsupervised learning setting, The standard RBF kernel function was used, $k(\mathbf{a},\mathbf{b}) = \exp(-\|\mathbf{a}-\mathbf{b}\|^2/(2\sigma^2))$ and set the lengthscale parameter σ to the median distance between all examples.

A visual comparison of the results in terms of activation maps for all methods is given in Fig. 5.3. They display the predictions given to each sample. The prediction maps show

Table 5.2: AUC results for Anomaly Detection images. The value for the best method for each image is in bold.

METHODS	RX	K-RX	RBIG	HYBRID
Cat-Island	*0.96*	*0.70*	**0.99**	**0.99**
WTC	**0.95**	*0.82*	**0.95**	**0.95**
Texas-Coast	**0.99**	*0.86*	*0.94*	**0.99**
GulfPort	*0.90*	**0.95**	**0.95**	**0.95**

a binary representation between change and non-change samples obtained from the model subject to a threshold. Results in all scenes demonstrate that (1) RX is a competitive method for detection, (2) KRX struggles to obtain reasonable results mainly due to the problem of hyperparameter tuning, (3) RBIG alone excels in all cases, while the hybrid approach (i.e. RX followed by RBIG) refines the results and yields clearer activation maps with sharper spatial detections.

Additionally, for a quantitative assessment of the results, it is customary to provide the ROC curves and to derive scores like the AUC from it. Figure 5.4 shows the ROC curves and Table 5.2 summarizes all AUC values for all images and methods. For each experiment, 1000 runs were performed for testing the significance of the methods based on the ROC profiles. The results are shown in Figure 5.5. Although the RBIG model achieves good results, RX model is able to compete and achieve results as good as RBIG for some images. The HYBRID model is able to keep the properties of the above mentioned models obtaining results equal or better than any other method. While KRX obtains a reasonable performance in some images, it clearly fails in some situations like the Cat-Island image. The low standard deviations show that all methods but the KRX are clearly robust with a little bit bigger standard deviation for the RX method in most cases.

5.3.3 Experiment 3: Real and Natural Changes

This section reports an experiment to analyze the performance of the proposed methods in change detection problems. The database is composed of different scenes with natural changes, whose characteristics are summarized in Table 5.1.

Data collection

Pairs of multispectral images were collected in such a way that they coincide at the same spatial resolution but at different acquisition time, the images are co-registered. The images are selected in such a way that an anomalous change happened between the two acquisition times. All the images were manually labeled finding the changed pixels. Labeling considered avoiding shadows, changes in lighting and natural changes

in vegetation which could compromise results evaluation. All images contain changes of different nature, which allows us to analyze and study how the algorithms perform in heterogeneous realistic scenarios. The Texas wildfire dataset is composed by a set of four images acquired by different sensors over Bastrop County, Texas (USA), and is composed by a Landsat 5 TM as the pre-event image and a Landsat 5 TM plus an EO-1 ALI and a Landsat 8 as post-event images. This phenomenon is considered the most destructive wildland-urban interface wildfire in Texas history and the interest region represent the 19.54%. The Argentina image represents an area burned between the months of July and August 2016 due to the high temperatures in these crop areas, the change region representing the 7.5% of the whole scene. The Chile dataset represents the Aculeo lake in central part of this country, which has now dried up completely. These images contrast the lake in 2014, when it still contained substantial water, and 2019, when it consisted of dried mud and green vegetation. Scientists attribute the lake's decline to an unusual decade-long drought, coupled with increased water consumption from a growing population, and the changed region represents a relevant 10.81% of the whole scene. The last dataset labeled as Australia shows the natural floods caused by Cyclone Debbie in Australia 2017. Storm damage resulted from both the high winds associated with the cyclone, and the very heavy rain that produced major riverline floods. The change samples represent an important portion of the scene, the 17.35% of pixels affected. Since our RBIG approach only takes the time t_1 image, these big changes do not have a critical impact on method's performance.

Numerical and Visual Comparison

Figure 5.6 shows the RGB composites of the pairs of images, the corresponding reference map and activation maps obtained. RBIG obtains clearly better results than the other methods in all cases; very good performance in three out of the four scenarios and a clear advantage in the most difficult one (Chile image). When dealing with highly skewed datasets, PR curves give a more informative picture of an algorithm's performance compared to ROC. Figure 5.7 shows both the ROC and the PR curves results for all methods and all the images. In all cases RBIG outperforms the other methods largely, thus suggesting the suitability of adopting a more direct approach of density estimation in the change detection problems too. A summary of the AUC values of all methods and scenarios is shown in Table 5.3. The RBIG approach is to able to estimate the change samples with a high accuracy overtaking in 7%, 3%, 6% and 5% respectively with respect the second best method.

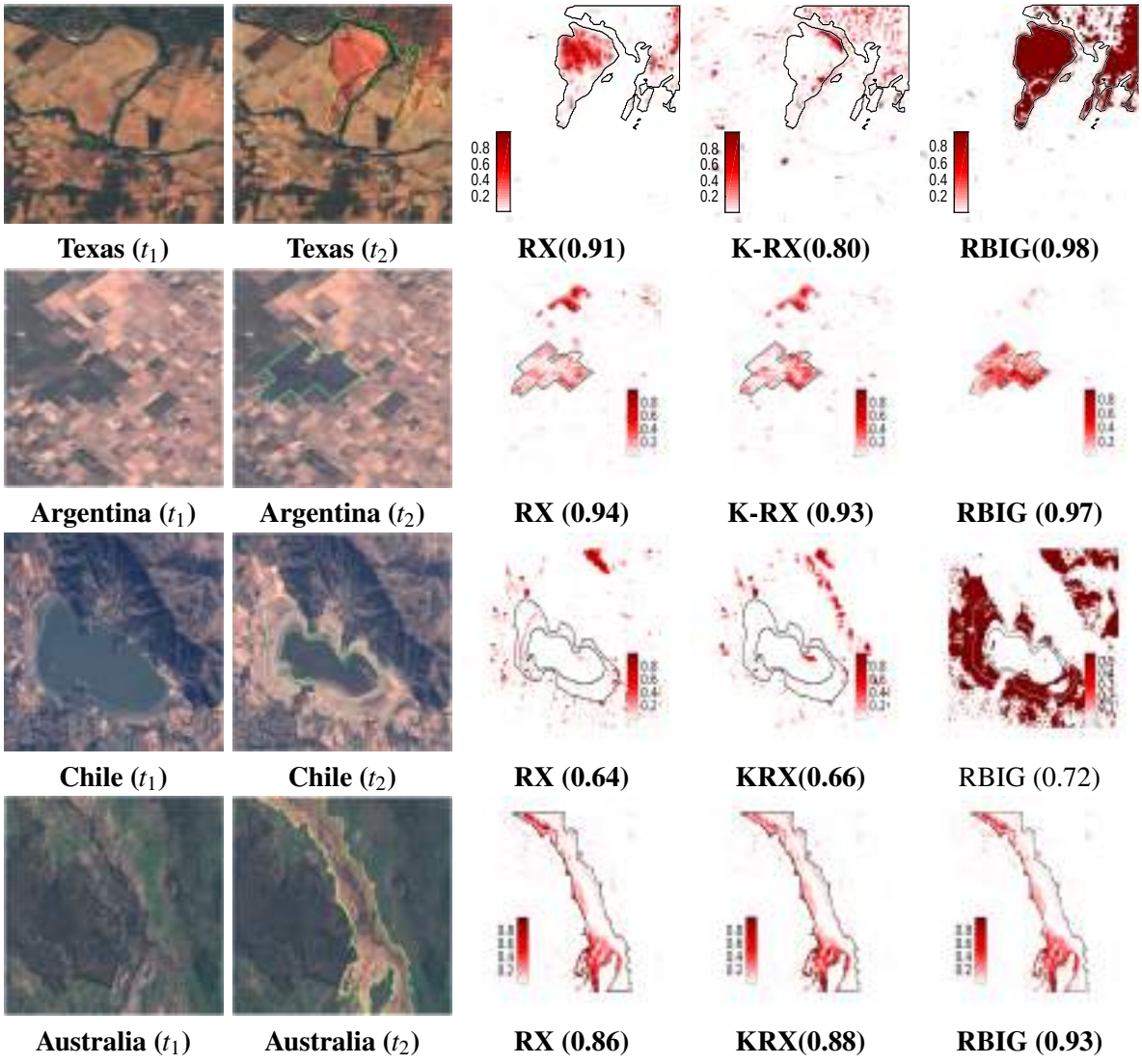


Figure 5.6: Change detection results for different images. First two columns show the images before and after the change, with the changed region highlighted in green. Columns three to five show the prediction maps for the different methods, the amount of change detected in each pixel is colored from white (less) to red (more). AUC values are given in parenthesis. The changed region is outlined in black to facilitate the visual inspection.

Table 5.3: AUC results for Change Detection images.
The best value for each image are in bold

METHODS	RX	K-RX	RBIG
Texas	*0.91*	*0.80*	**0.98**
Argentina	*0.94*	*0.93*	**0.97**
Chile	*0.64*	*0.66*	**0.72**
Australia	*0.86*	*0.88*	**0.93**

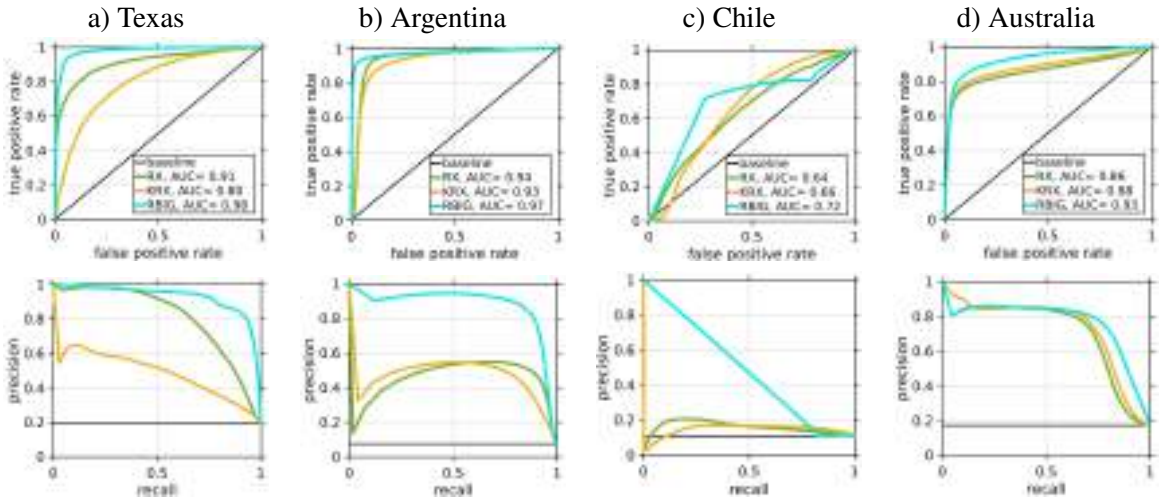
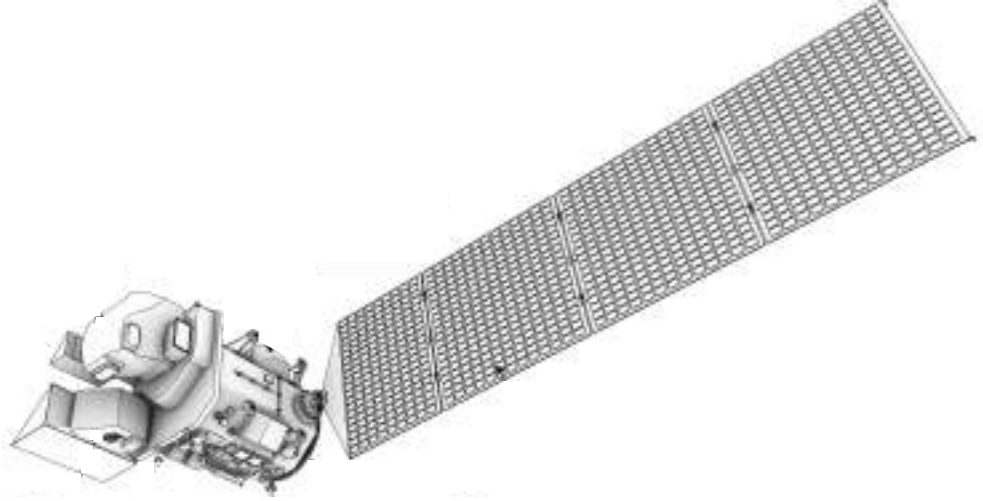


Figure 5.7: ROC (top row) and Precision-Recall (bottom row) curves for change detection problems.

5.4 Specific contributions

In this chapter, a novel detector was introduced to cope with anomaly and change detection problems in remote sensing image processing. The method is based in a unsupervised setting with no parameters to fit. The model assumption is based on detecting anomalies by estimating probabilities of pixels. The proposed methods are based on visual inspection (activation maps) and accuracy values (AUC). The algorithms testing is implemented in a wide range of remote sensing images, in a diversity of problems, dimensionality and number of examples. In addition, a hybrid approach is applied after a regular anomaly detector: this facilitates the density estimation and improves the results notably.



6. CONCLUSIONS

This Thesis proposed novel machine learning algorithms for the detection of anomalies in Remote Sensing imagery. Several methods were developed and tested under the kernel methods for improve accuracy, versatility and computational efficiency. On the other hand, a novel proposal was presented based on explicit PDF estimation under the rotation-based iterative Gaussianization framework. In summary, several methods were developed for detection of anomalies and changes in satellite images.

On the efficiency of the kernel Cook's distance for ACD

The kernel Cook's distance for anomalous change detection settings was focused on remote sensing image change detection problems. The key in the proposed methodology was to redefine the anomalous change detection problem in a reproducing kernel Hilbert space where the data are mapped to. This endorses the methods with improved capacity and flexibility since nonlinear feature relations (and hence outliers) can be better identified. However, the obtained kernelized method encounters huge computational problems in practice, which hampers its applicability and wider adoption. To resolve this problem, computationally efficient techniques were proposed based on random Fourier features and low-rank Nyström approximations, and compared their capabilities in a wide range of both simulated and real changes. The Nyström approximation excelled over the rest of the implementations, in both simulated and real scenarios, and in terms of accuracy and efficiency. Future work will study other related kernel diagnostic measures. Extension to online and multi-change problems are also topics of further research.

On the Kernel RX distance for ACD

The family of kernel-based anomaly change detection algorithms was extended to the standard methods like the RX detector (Lu et al., 1997; Reed & Yu, 1990b), and many others in the literature (Theiler & Perkins, 2006; Theiler et al., 2010). The key in the proposed methodology was to redefine the anomaly detection in a reproducing kernel Hilbert space where the data are mapped to. This endorses the methods with improved capacity and flexibility since nonlinear feature relations (and hence anomalies) can be identified. The introduced methods generalize the previous ones since they account for higher-order dependencies between features. The proposed methods obtain better results than their linear counterpart for all the performed experiments. Implementations of the methods were provided and a database of pairs of images with anomalous changes that can be found in real scenarios.

In practical terms, kernel ACD methods presented here yielded improved results over their linear counterparts in multiple situations. The robustness of this conclusion performing experiments was tested in a wide range of problems. Experiments with different complexity levels were designed : synthetic anomaly, real but manually introduced anomaly and real data where the anomaly has been manually labeled. The performance in data coming from different sensor (multi and hyperspectral) was analyzed, showing that the kernel methods are robust to different number of input data dimensions as expected (Gómez-Chova et al., 2011). Standard metrics (AUC and detection) were adopted and the results over several runs were averaged to avoid skewed conclusions.

Interestingly, the EC assumption may be still valid in Hilbert spaces, especially when high pervasive distortions mask anomalous targets. This observation opens the door to the study of the anomalies distribution in Hilbert spaces in the future. A second important conclusion of this approach to be highlighted is that, although the XY family seams to work better for the K-EC-ACD method, among all 16 methods implemented, there was no a clear winner between all methods. After all, each problem has its own characteristics and the different methods adapt to different particularities. In the future, the plan is to extend the study with low-rank, sparse and scalable kernel versions to cope with high computational requirements.

On the efficiency of the nonlinear RX anomaly detectors

The family of efficient nonlinear anomaly detection algorithms based on the RX method was developed to cope with anomaly detection approach. The theory of reproducing kernels was proposed as well as several efficient methods to approximate the kernel one. The

kernel Reed-Xiaoli (KRX) detector was improved using efficient and fast techniques based on feature maps and low-rank approximations. Among all methods, both the Nyström and the equivalent low-rank (LRX) approximation achieves the best results and yields a more efficient and accurate non-linear RX method to be applied in practice. For future research, we plan to study the behaviour of fast approximations for alternative KRX variants (Theiler & Grosklos, 2016; Theiler & Grosklos, 2016). Note that the presented methodologies for fast KRX can be applicable to other kernel anomaly detectors, in local settings, and for real-time detection.

On the rotation-based iterative Gaussanization

A novel detector based on multivariate Gaussianization was proposed. The methodology copes with anomaly and change detection problems in remote sensing image processing, and meets all requirements of the problems: is an unsupervised method with no parameters to fit, it can deal with large amount of data, and it is more accurate to competing approaches. The model assumption is based on detecting anomalies by estimating probabilities of pixels. The proposed method excelled quantitatively (AUC, ROC and PR curves) and qualitative based on visual inspection over the rest of the implementations, in both anomaly and change detection. The evaluation considered a wide range of remote sensing images, in a diversity of problems, dimensionality and number of examples. Also, a hybrid approach was suggested where the Gaussianization method is applied after a regular anomaly detector: this facilitates the density estimation and improves the results notably.

Related works that support this thesis.

The Thesis is completed by an annex which includes a compendium of peer-reviewed publications in remote sensing international journals, summarized as follows:

1. *Kernel Anomalous Change Detection for Remote Sensing Imagery*. Padrón-Hidalgo, J. A. and Laparra, V. and Longbotham, N and Camps-Valls, G. IEEE Transactions on Geoscience and Remote Sensing 10, vol 57, pages: 7743-7755, 2019. Journal Impact Factor (5.85). Q1: Electrical and Electronic Engineering. Q1: Remote Sensing.
2. *Efficient Nonlinear RX Anomaly Detectors*. José A. Padrón Hidalgo and Adrián Pérez-Suay and Fatih Nar and Gustau Camps-Valls IEEE Geoscience and Remote Sensing Letters, pages: 1-5, 2020. Journal Impact Factor (3.83). Q1: Electrical and Electronic Engineering. Q1: Geochemistry and Geophysical.
3. *Efficient Kernel Cook's Distance for Remote Sensing Anomalous Change Detection*. Padrón-Hidalgo, J.A. and Pérez-Suay, A. and Nar, F. and Laparra, V. and Camps-

Valls, G. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol 13, pages: 5480 - 5488, 2020. Journal Impact Factor (3.83). Q1: Electrical and Electronic Engineering. Q1: Geographic Physical.

4. *Unsupervised Anomaly and Change Detection with Multivariate Gaussianization*. Padron, J. and Laparra, V. and Camps-Valls, G. Submitted to IEEE Transactions on Geoscience and Remote Sensing, 2020. Journal Impact Factor (5.85). Q1: Electrical and Electronic Engineering. Q1: Remote Sensing.

Other related publications in conferences and workshops are listed here too for completeness:

1. *Kernel Anomalous Change Detection*. Jose A. Padrón Hidalgo and Valero Laparra and Gustau Camps-Valls Young Professionals Conference on Remote Sensing , Aachen, Germany 2018
2. *Nonlinear Cook Distance for Anomalous Change Detection*. Jose A. Padrón Hidalgo and Adrián Pérez-Suay and Fatih Nar and Gustau Camps-Valls 2018 IEEE International Geoscience and Remote Sensing Symposium, València, Spain 2018
3. *Randomized RX for Target Detection*. Fatih Nar and Adrian Perez-Suay and Jose Antonio Padron and Gustau Camps-Valls 2018 IEEE International Geoscience and Remote Sensing Symposium, València, Spain 2018

Acknowledgments

The research activities were supported by the European Research Council (ERC) under the ERC-CoG-2014 SEDAL project (grant agreement 647423) and the Spanish Ministry of Economy, Industry and Competitiveness under the 'Network of Excellence' program (grant code TEC2016-81900-REDT). In addition, Jose A. Padrón was supported by the Grisolia grant from Generalitat Valenciana (GVA) with code GRISOLIA/2016/100.

SUMMARY IN SPANISH

La Tesis se basa en un compendio de publicaciones de nivel científico publicadas en revistas de reconocimiento internacional. Todas estas publicaciones se centran en el desarrollo de modelos de aprendizaje automáticos para la detección de cambios anómalos entre pares de imágenes, así como la detección de anomalías en imágenes de teledetección. Las publiaciones en formato original de cada revista se localizan en el apartado de anexos. A continuación se expone un resumen de esta Tesis en castellano con el objetivo que pueda llegar a los usuarios de la Universitat de València, especialmente como inspiración aquellos estudiantes que se inician en este interesante campo del aprendizaje automático en la Teledetección.

Motivación y objetivos

La Tierra es un sistema complejo de redes dinámicas y en los últimos cientos de años la actividad humana ha precipitado enormes cambios en el Planeta. No hace falta decir que en la actualidad el desafío más importante para la ciencia es detectar y determinar las causas de tales cambios. En este escenario, los datos de observación de la Tierra nos permiten detectar automáticamente anomalías en la cubierta terrestre tanto en el dominio espacial como en el temporal. Esto es posible actualmente mediante el uso de imágenes satelitales de alta resolución y de series temporales de imágenes, junto con poderosas técnicas estadísticas para procesarlas. Sin embargo, en los últimos años, los grandes y heterogéneos flujos de datos adquiridos por las constelaciones de satélites, obstaculizan la adopción de técnicas estadísticas avanzadas de aprendizaje automático para la detección tanto de anomalías como de cambios anómalos entre imágenes de satélites. El objetivo principal de esta Tesis es desarrollar y aplicar detectores novedosos y robustos para detectar aquellos eventos o situaciones que se consideran atípicos o fuera de lo normal como es el caso de las sequías, inundaciones, incendios forestales, urbanizaciones y otros ejemplos que a menudo suelen aparecer en la monitorización de la Tierra. En la actualidad la mayoría de los algoritmos que tratan la detección de anomalías y cambios anómalos suelen ser cuestionados por la precisión o la eficiencia a la hora de detectar dichos eventos. Esta Tesis se basa en dos marcos principales para desarrollar, mejorar e implementar detectores robustos. Por un

lado los métodos basados en el Kernel proporcionan un marco teórico consistente y bien fundamentado para el desarrollo de técnicas que permiten lidiar con la no linealidad de los datos y presentan propiedades útiles cuando se trata de un número bajo de muestras de entrenamiento en datos de alta dimensionalidad. Uno de los problemas a los que nos enfrentamos con estos métodos es el alto coste computacional debido al gran tamaño que presentan las imágenes satelitales. De aquí se deriva otro de los objetivos de esta Tesis que es desarrollar modelos automáticos, rápidos y eficientes basados en aproximaciones del Kernel. El objetivo es que estos métodos superen en precisión de detección a los métodos lineales. Por otra parte, otro de los marcos utilizados se basa en la estimación explícita de la densidad. Este objetivo se centra en la necesidad de desarrollar algoritmos de detección que se entrenen de manera no supervisada, ya que los métodos basados en Kernel y sus aproximaciones necesitan ajustar de forma manual o mediante la validación cruzada sus parámetros. Esta técnica que se basa en la Gaussianización multivariante, permite estimar con precisión densidades multivariantes, un problema clásico en estadística y el aprendizaje automático sobre todo cuando los datos tienen una gran dimensionalidad. A su vez, este método es empleado de manera no supervisado para la detección de cambios y anomalías en las imágenes de teledetección.

Metodología

La Tesis aborda problemas relacionados con la detección de cambios anómalos que implícitamente involucra la detección de cambios y la detección de anomalías como casos particulares, diseñando algoritmos de aprendizaje automático que resuelvan estos problemas. Estudiamos el rendimiento de todos los algoritmos propuestos en un número representativo de imágenes satelitales multiespectrales y de alta resolución espacial como AVIRIS, Sentinel-2, WorldView-2, MODIS, Quickbird y Landsat8, así como en una amplia gama de situaciones relacionadas con sequías, incendios forestales, inundaciones y urbanización. Los métodos propuestos se basan principalmente en la estimación de distancias y probabilidades. En el caso de los modelos basados en distancia se centraron en el conocido detector Reed-Xiaoli (RX) y su familia de detectores, así como en la distancia de Cook y sus aproximaciones. Ambos enfoques hacen referencia a versiones lineales y no lineales para la detección de anomalías y cambios anómalos en imágenes de teledetección. La familia de los métodos RX es extendida a su versión no lineal mediante el uso de kernels de forma que es capaz de mejorar la precisión de la detección con respecto a los métodos lineales originales. Por otra parte, la distancia de Cook se extiende mediante el uso de kernels para abordar los problemas de cambios anómalos. Además, se

utilizan aproximaciones del Kernel basadas en el método de características aleatorias de Fourier y el método de Nyström que ayudan a mejorar la eficiencia, el coste computacional y la precisión de los modelos. En el caso de los modelos basados en estimación de probabilidades se ha utilizado la metodología de Gaussianization multivariante iterativa que permite describir cualquier distribución multivariante y hace un uso eficiente de los recursos de memoria y computación. Es un método no supervisado que no necesita ajustar ningún parámetro. Se demostró la eficiencia del método en experimentos que implican tanto la detección de anomalías como la detección de cambios en diferentes conjuntos de imágenes de satélites. Para la detección de anomalías proponemos dos enfoques. El primero utilizando directamente la Gaussianización iterativa basada en rotación (RBIG) y el segundo utilizando un modelo híbrido que combina la Gaussianización y el método Reed-Xiaoli (RX) que habitualmente es utilizado en la detección de anomalías.

Métodos Kernel para la detección de anomalías y cambios anómalos.

El marco teórico de aprendizaje mediante los métodos del kernels, han surgido como uno de los escenarios más apropiados para el análisis de datos de teledetección en la última década. Los métodos kernels permiten generalizar los algoritmos expresandos en términos de su matriz de Gram, de manera que se tengan en cuenta las relaciones de características de orden superior (no lineales), pero aun así trabajando mediante álgebra lineal. Los métodos Kernel destacan en el tratamiento de datos con tamaños que van de bajos a moderados, pueden acomodar datos de múltiples fuentes, modelar distribuciones complejas con funciones kernel flexibles, y hacer frente a datos de alta dimensionalidad. Además se ajustan adecuadamente a las características particulares de las señales de Observación de la Tierra, tales como series temporales muestreadas de manera desigual, datos faltantes, distribuciones no gaussianas y procesos no estacionarios. Los métodos Kernel han sido tradicionalmente diseñados para problemas de clasificación y regresión. Sin embargo, la familia de métodos Kernel se expande actualmente a la detección de cambios multitemporales, la estimación de dependencia no lineal, la prueba de hipótesis, y la detección de anomalías, que constituyen el eje central de esta Tesis.

A partir de la ventaja que ofrece la formulación basada en métodos kernels se han extendido dos métodos lineales altamente utilizados para la detección de anomalias y cambios anómalos. Por un lado, se ha desarrollado la kernelización de la distancia Cook. Esta distancia es usada para la detección de cambios anómalos en un esquema donde el indicador de anomalías proviene de la evaluación estadística de los residuos de un regresor entre imágenes en adquisiciones de tiempos diferentes. En particular se desarrolló la

formulación matemática para sustituir las regresión lineal por el método Kernel Ridge Regression. Por otro lado se introducen métodos Kernel para implementar una extensión no lineal (KRX) de la familia de detectores de cambios anómalos basados en RX. En particular, se centró en los algoritmos que utilizan la distribución de contorno gaussiano y elíptico y los se extiendes a sus equivalentes no lineales basados en la teoría de la reproducción del Kernel en el espacio de Hilbert. Se ilustra el rendimiento de los métodos introducidos en una amplia serie de imágenes de distintos satélites.

Cabe destacar que ambas propuestas presentan un alto coste computacional debido al trabajo con imágenes de satélites. Para solucionar este problema se han utilizado distintas técnicas para obtener aproximaciones eficientes de la función kernel. Una de las técnicas se basa en el uso de bases aleatorias de Fourier (Rahimi & Recht, 2007). Estas bases definen un mapeo que toma los datos en el espacio de entrada y los transfiere a un nuevo espacio euclideo de dimensiones finitas, donde el problema es linealmente separable y el producto interno de los datos mapeados se aproxima a la función Kernel. Así, el algoritmo proporcionado es computacionalmente más eficiente y como se mostrará a lo largo de esta Tesis, converge a velocidades similares y a escalas de error similares. Otra técnica se basa en el uso de las características aleatorias ortogonales (ORF), las cuales son similares a la técnica anterior pero imponiendo ortogonalidad sobre la matriz de transformación lineal. Además, se han considerado aproximaciones de bajo rango (Fine & Scheinberg, 2001) de la matriz Kernel como por el ejemplo el método de Nyström, que permiten reducir las complejidades del tiempo de ejecución a la hora de invertir la matriz kernel. Todas estas aproximaciones se implementan y testean a lo largo de esta Tesis tanto para la detección de anomalías como para la detección de cambios anómalos.

Estimación de densidad con transformación gaussiana

La Gaussianización iterativa basada en rotación (RBIG) es un método no paramétrico para la estimación de la densidad de distribuciones multivariadas. Se utilizó el método RBIG como no aprendizaje supervisado para detectar anomalías y cambios en las imágenes de teledetección. La metodología de Gaussianización permite estimar con precisión las densidades multivariantes, un problema clásico en estadística y el aprendizaje automático. El método RBIG se fundamenta en la idea de la Gaussianización multivariada, que consiste en buscar una transformación que convierta un conjunto de datos multivariados a un dominio en el que los datos mapeados sigan una distribución normal multivariada. Por lo tanto, aplicando la fórmula del cambio de distribucion bajo transformaciones, el modelo permite estimar la probabilidad en cualquier punto del dominio original. En nuestro caso

se utilizó esta estimación para determinar que los píxeles de baja probabilidad estimada se consideran anomalías. Esta es la misma definición de anomalia que la usada por el método RX (mencionado arriba) el cual asume distribución Gaussiana en el dominio original.

Un aspecto importante a tener en cuenta son las características intrínsecas de los datos utilizados para estimar la densidad, lo que tiene implicaciones en la calidad de la estimación. Cuando la distribución contiene incluso un número moderado de anomalías, una estimación precisa de la densidad arrojará las anomalías como puntos regulares, es decir, no anómalos. Esto depende en gran medida de la flexibilidad de el tipo de modelo utilizado. Cuando el modelo es rígido como en el caso del RX, esto no es un problema, ya que no puede adaptarse a las anomalías. En el caso del KRX se puede controlar este efecto ajustando sus parámetros principales incluyendo el término de regularización pero requiere datos etiquetados. Este es un aspecto importante a tener en cuenta sobre todo en el escenario de detección de anomalías, donde todos los datos (incluidas las muestras anómalas) se utilizan para estimar la densidad. Por lo tanto, proponemos utilizar un modelo híbrido que combina el modelo RX (demasiado rígido) con el modelo RBIG (demasiado flexible). El modelo híbrido primero selecciona los datos con mayor probabilidad de no ser anómalos utilizando el método RX y luego utiliza estos datos para aprender la transformación de Gaussianización con el modelo RBIG. Esto trata de evitar el uso de datos anómalos para entrenar a RBIG, que después de todo está destinado a aprender el fondo o la distribución de datos no anómalos. El número de puntos seleccionados como no anómalos en el primer paso definirá el equilibrio entre flexibilidad y rigidez. Por otra parte, se aplicó la misma teoría para hacer frente a los problemas de detección de cambios en teledetección. Es importante señalar que, en este caso, los datos utilizados para estimar la densidad de probabilidad (primera imagen) no contienen anomalías, por lo que no es necesario el modelo híbrido en el problema de detección de cambios.

Conclusiones

Basados en el aprendizaje automatizado se ha desarrollado una variedad de modelos tanto para la detección de anomalías como para la detección de cambios anómalos. Se implementaron detectores novedosos y robustos capaces de detectar las anomalías con precisión y eficiencia a partir de diferentes marcos teóricos. Por una parte, los métodos basados en distancias se fundamentaron bajo la teoría del Kernel, así como sus aproximaciones eficientes para reducir el costo computacional. Por otro lado, se desarrollaron métodos que son estimadores de probabilidad basados en la Gaussianización multivariante iterativa para la detección de anomalías y cambios en imágenes de teledetección. Todos estos

detectores se implementaron y caracterizaron en distintos escenarios tanto simulados como en situaciones reales, detectando anomalías tales como: sequías, incendios forestales, inundaciones y urbanización, a partir de los sensores AVIRIS, Sentinel-2, WorldView-2, MODIS, Quickbird y Landsat8. Todas estas imágenes fueron etiquetadas manualmente debido al limitado acceso a las bases de datos para la validación de los mismos. Cabe resaltar que estas bases de datos quedaran a la disposición de la comunidad científica así como los códigos de todos los métodos empleados en esta Tesis. Los capítulos fueron el resultado de diferentes inestigaciones científicas, los cuales muestran la teoría de los métodos implementados así como las distintas aplicaciones en la observación de la Tierra.

En el capítulo 1 se hizo un breve recorrido del importante uso de la teledetección en la observación de la Tierra. Se abordaron los principales conceptos de detección de anomalías, detección de cambios y detección de cambios anómalos en el contexto de la teledetección. Además, se hizo una breve reseña de los modelos de aprendizaje automáticos que más se utilizan en la bibliografía haciendo énfasis en los que se han implementado.

En el capítulo 2 se presentó una versión kernelizada de la distancia de Cook (Cook, 1977). Este método que anteriormente había sido usado para la detección de anomalías en datos estadísticos fue desarrollado para detección de cambios anómalos en imágenes de teledetección. La clave de la metodología propuesta radicó en desarrollar la versión Kernel de la distancia de Cook debido a la carencia de la versión linear de detectar anomalías en datos con distribuciones no lineales. Por lo tanto, mapear los datos hacia el espacio de Hilbert respaldó con una mayor capacidad y flexibilidad estos métodos, ya que las relaciones de características no lineales (y por lo tanto los valores atípicos) pudieron identificarse con mayor efectividad. Sin embargo, el método propuesto enfrentó problemas de implementación debido al coste computacional, lo que dificultó su aplicabilidad y su adopción en la práctica. Para resolver este problema, se propusieron técnicas computacionales eficientes basadas en características aleatorias de Fourier y las aproximaciones de Nyström de bajo rango. Se compararon las capacidades de los métodos en una amplia gama de cambios anómalos tanto simulados como reales. La aproximación de Nyström sobresalió sobre el resto de las implementaciones, tanto en los escenarios simulados como en los reales, y en términos de precisión y eficiencia. En el futuro se propone el estudio de otras medidas de diagnóstico relacionadas con el Kernel. Por otra lado, enfocarlos a los problemas en línea y además, a los cambios múltiples.

En el capítulo 3 se implementaron modelos enfocados a la detección de cambios anómalos. Esta vez, se desarrolló y se amplió la familia de algorítmos de detección basados en los detectores RX (Reed & Yu, 1990b), y sus versiones elípticas (Theiler et al.,

2010) haciendo uso de la teoría del Kernel. La clave de la metodología propuesta fue redefinir la detección de cambios en un espacio infinito (Hilbert) donde los datos son mapeados e implementar la versión Kernel de su contraparte lineal (KRX). Esto ayuda a los métodos a identificar las anomalías con mayor facilidad. Los métodos introducidos generalizan las versiones lineales anteriores, ya que tienen en cuenta las dependencias de orden superior entre las características. Los métodos propuestos obtuvieron mejores resultados que su contraparte lineal en todos los experimentos realizados. Se comprobó mediante experimentos en una amplia gama de problemas la solidez de esta conclusión. Se diseñaron experimentos con diferentes niveles de complejidad: anomalías sintéticas, anomalías reales pero introducidas manualmente y datos reales donde la anomalía ha sido etiquetada manualmente. Se analizó el rendimiento en los datos procedentes de diferentes sensores (multi e hiper espectrales) mostrando que los métodos Kernel son robustos para diferentes dimensiones de datos de entrada como se esperaba. Curiosamente, la asunción del contorneado elíptico puede seguir siendo válida en los espacios de Hilbert, especialmente cuando las distorsiones de alta penetración enmascaran objetivos anómalos. Esta observación abre la puerta al estudio de la distribución de las anomalías en los espacios Hilbert en el futuro. Una segunda conclusión importante de este estudio que hay que destacar es que, aunque la familia XY parece funcionar mejor para el método K-EC-ACD, entre los 16 métodos aplicados, no se observó un claro ganador en todos los métodos. Después de todo, cada problema tiene sus propias características y los diferentes métodos se adaptan a las diferentes particularidades. En el futuro se pretende ampliar el estudio con versiones del Kernel de bajo rango, dispersas y escalables para hacer frente a los altos requisitos de coste computacional.

Hasta el momento, el trabajo ha estado dirigido a la detección de cambios anómalos entre pares de imágenes. En el capítulo anterior se propuso la implementación del KRX, y en el capítulo 4 se propusieron varios métodos que aproximaran al KRX enfocados a la detección de anomalías en una sola imagen. La versión kernelizada del método RX fue mejorada utilizando técnicas eficientes y rápidas basadas en las caracterísicas aleatorias de Fourier, las características aleatorias ortogonales, las aproximaciones de bajo rango incluyendo en esta categoría el método de Nyström. Entre todos los métodos, tanto la aproximación de Nyström como la equivalente de bajo rango (LRX) lograron los mejores resultados y crearon un método no lineal más eficiente y preciso para ser aplicado en la práctica. Para futuras investigaciones, se pretende estudiar el comportamiento de las aproximaciones eficientes para otras variantes alternativas del KRX.

Una vez finalizado un estudio profundo basado en la teoría del Kernel, se ha llegado a

la conclusión que dichos modelos son costosos computacionalmente comparados con su contraparte lineal. Estos modelos también presentan parámetros que deben ser ajustados para poder obtener la mayor eficiencia. En nuestro caso, estos parámetros fueron ajustados usando la validación cruzada, lo que significa que se desarrolla de una manera supervisada maximizando el área bajo las curvas ROC. Esto permitió desarrollar en el capítulo 5 un novedoso detector basado en la Gaussianización multivariante iterativa. La metodología hace frente tanto a los problemas de detección de anomalías como a la detección de cambios en el procesamiento de imágenes de teledetección. Por lo que cumple todos los requisitos de los problemas antes mencionados: es un método no supervisado sin parámetros a ajustar, puede tratar con una gran cantidad de datos, y es más preciso para los enfoques que compiten entre sí. El modelo se basa en la detección de anomalías mediante la estimación de las probabilidades de los píxeles. El método propuesto sobresalió cuantitativamente teniendo en cuenta valores de AUC proporcionadas por las curvas ROC y las curvas de precisión. Además, cualitativamente fue superior teniendo en cuenta la inspección visual sobre el resto de las implementaciones, tanto en la detección de anomalías como de cambios entre imágenes. En la evaluación se consideró una amplia gama de imágenes de teledetección, en una diversidad de problemas, dimensionalidad y número de ejemplos. También se sugirió un enfoque híbrido en el que se aplicó el método de Gaussianización después de un detector de anomalías regular, lo que facilitó la estimación de la densidad y mejoró notablemente los resultados.

En general se puede concluir que los algoritmos desarrollados basados en el Kernel, sus aproximaciones y además los modelos basados en estimar densidades de probabilidades pueden ser implementados y puestos en práctica, ya que son capaces de detectar con muy buena precisión las anomalías en diferentes situaciones reales. Se demostró a través de los diferentes experimentos tantos sintéticos como reales que los métodos son robustos, en las diversas situaciones de la vida real y en relación a las características de las imágenes teniendo en cuenta diferentes resoluciones tanto espaciales como espectral. Estos modelos ayudarán al monitoreo de las zonas de difícil acceso reduciendo así en gran proporción el coste económico que estos pueden causar.

Publications

The achievements and conclusions of this work have been published in the following papers in high quality international journals:

Publication I

J. A. Padrón-Hidalgo, V. Laparra, N. Longbotham and G. Camps-Valls, "Kernel Anomalous Change Detection for Remote Sensing Imagery," in *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 10, pp. 7743-7755, Oct. 2019, doi:10.1109/TGRS.2019.2916212.

Publication II

J. A. Padrón-Hidalgo, A. Pérez-Suay, F. Nar and G. Camps-Valls, "Efficient Non-linear RX Anomaly Detectors," in *IEEE Geoscience and Remote Sensing Letters*, pp. 1-5, 2020. doi: 10.1109/LGRS.2020.2970582.

Publication III

J. A. Padrón-Hidalgo, A. Pérez-Suay, F. Nar, V. Laparra and G. Camps-Valls, "Efficient Kernel Cook's Distance for Remote Sensing Anomalous Change Detection," in *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 13, pp. 5480-5488, 2020, doi: 10.1109/JSTARS.2020.3020913.

Publication IV

Padrón-Hidalgo, J. A. and Laparra, V. and Camps-Valls, G. "Unsupervised Anomaly and Change Detection with Multivariate Gaussianization", Submitted to *IEEE Transactions on Geoscience and Remote Sensing*, 2020.

Efficient Kernel Cook's Distance for Remote Sensing Anomalous Change Detection

José Antonio Padŕon-Hidalgo, Adrián Pérez-Suay, Fatih Nar, Valero Laparra, and Gustau Camps-Valls

***Abstract*—Detecting anomalous changes in remote sensing images is a challenging problem, where many approaches and techniques have been presented so far. We rely on the standard field of multivariate statistics of diagnostic measures, which are concerned about the characterization of distributions, detection of anomalies, extreme events, and changes. One useful tool to detect multivariate anomalies is the celebrated Cook's distance. Instead of assuming a linear relationship, we present a novel kernelized version of the Cook's distance to address anomalous change detection in remote sensing images. Due to the large computational burden involved in the direct kernelization, and the lack of out-of-sample formulas, we introduce and compare both random Fourier features and Nyström implementations to approximate the solution. We study the kernel Cook's distance for anomalous change detection in a *chronochrome* scheme, where the anomalousness indicator comes from evaluating the *statistical leverage* of the residuals of regressors between time acquisitions. We illustrate the performance of all algorithms in a representative number of multispectral and very high resolution satellite images involving changes due to droughts, urbanization, wildfires, and floods. Very good results and computational efficiency confirm the validity of the approach.**

***Index Terms*—Anomalous change detection (ACD), Cook's distance, efficiency, influential points, kernel methods, Nyström method, random Fourier features, statistical leverage.**

I. Introduction

THE Earth's surface is constantly changing due to natural events and various anthropogenic interventions. Natural events can be repetitive ones such as seasonal changes as well as extreme or rare events such as disasters. Newly constructed man-made structures, urbanization, and agriculture activities can be given as examples of anthropogenic interventions [1]. However, observing such changes in a timely and accurate manner is challenging since the Earths' surface is very large and complex, and changes are constantly happening, which may be pervasive or anomalous. Change detection (CD) using remote sensing (RS) images is an active field of research with many systematic methods and procedures to capture the changes on the earth surface. CD methods that are applied to RS images can be acquired from satellite or airborne platforms [2].

CD is extremely important because it allows us to improve predictions and our understanding of events occurring over the entire surface of the earth, such as floods and droughts [3], [4], using Landsat 7 images. In addition, the developed CD methods can help us improve designing and implementing urban monitoring [5]. However, factors such as seasonal differences, atmospheric effects, sensor noise, and registration errors create spurious changes that decrease the performance of the CD methods [1]. In addition, the Earth's surface is complex and heterogeneous, while obtained images can be multimodal or multisource with different spectral and temporal resolutions, which further increase the complexity of the development of robust, accurate, and fast CD methods. The recent advances in RS sensors, statistical models, and computational power, as well as an immense amount of data availability, have provided additional possibilities and challenges in RS image processing [6]–[8]. Most importantly, the increasing spatial and temporal resolution of globally available satellites such as Sentinel-2 gives a unique opportunity to monitor regular and extreme events. In addition, the use of very high-resolution satellite imagery (such as Digital Globe QuickBird) is becoming increasingly important for RS applications.

A related field to CD called anomalous change detection (ACD) is concerned about a slightly different problem [9]: the ACD setting differs from standard CD because the objective is to identify only rare (or anomalous) events, ignoring the regular (or pervasive) ones. These pervasive differences may be due to calibration, illumination, look angle, and even the choice of RS satellite. By contrast, the anomalous changes are assumed to be relatively rare and can be highlighted in a minor part of the image. Although there were related studies before, first focused study of ACD was proposed by Theiler and Perkins using a machine learning approach [9] with many other subsequent studies of Theiler and his colleagues [10]–[17]. In the literature, ACD was tackled using distribution-based [9], [12], distance-based [13], classifier-based [11], and reconstruction-based [18] approaches. Note that ACD is also closely related to anomaly detection [16] and novelty detection [19], where all employ similar approaches. However, most ACD methods

Manuscript received May 14, 2020; revised August 10, 2020; accepted August 23, 2020. Date of publication September 1, 2020; date of current version September 25, 2020. This work was supported in part by the European Research Council (ERC) through the ERC-CoG-2014 SEDAL Project under Grant 647423 and in part by the Spanish Ministry of Economy, Industry and Competitiveness through the "Network of Excellence" Program under Grant TEC2016-81900-REDT. The work of José Antonio Padrón-Hidalgo was supported by Generalitat Valenciana under Grisolia Grant GRISOLIA/2016/100. *(Corresponding author: José Antonio Padŕon-Hidalgo.)*

José Antonio Padŕon-Hidalgo, Adrián Pérez-Suay, Valero Laparra, and Gustau Camps-Valls are with the Image Processing Laboratory, Universitat de València, 46980 Valencia, Spain (e-mail: joseantoniopadronhidalgo@gmail.com; gustau.camps@uv.es).

Fatih Nar is with the Ankara Yıldırım Beyazıt University, Ankara 06010, Turkey.

This article has supplementary downloadable material available at https://ieeexplore.ieee.org, provided by the authors.

Digital Object Identifier 10.1109/JSTARS.2020.3020913

are linear, which limit their success and applicability to the real-world problems [20]. Researchers introduced nonlinearity to overcome this limitation, i.e., using neural networks [21] and kernel methods [20], [22], [23]. Although kernel methods bring excellent performance, they are computationally demanding, so efficient approximations are needed [22]. Interested readers can read more about ACD in [10], [17], [24], and [25] that are, respectively, comparison, analysis, review, and tutorial style studies.

The ACD settings can be properly framed in statistical terms. However, the concept of *anomalousness* is elusive and difficult to define concretely. Nevertheless, identifying influential points in multivariate data distributions is an active field of research in statistics, information theory, and machine learning. Main applications involve characterizing distributions, detecting anomalies, extremes and changes, and assessing robustness [14], [15]. Detection of such influential points also have relevant applied implications for climate, health, and social sciences, and in a wide diversity of engineering and computer science problems. It is important to remark that we aim to detect anomalous (extreme) changes, i.e., not pervasive changes related to, for example, illumination conditions. Therefore, we will refer to anomalies among two images, leverage points, or changes interchangeably.

The interest to find anomalous changes is very broad, and many methods have been proposed in the literature, ranging from equalization-based approaches that rely on whitening principles [26] to multivariate methods that extract distinct features out of the change (difference) image [27] and that reinforce directions in feature spaces associated with noisy or rare events [28], [29], as well as regression-based approaches such as in the chronochrome [30], [31], where a regression model approximates the next incoming image and big residuals are associated with anomalies. In this article, we build our nonlinear ACD method on this latter chronochrome approach based on Cook's distance [32], where our initial efforts can be seen in [22]. Among other measures, we preferred Cook's distance since it allows robust fitting despite data are being contaminated by outliers (or anomalies). Many diagnostic measures have been introduced other than the seminal work of Cook such as linear regression [33], [34], penalized (ridge) regression [35], sparse regression models like LASSO [36] as *parametric models*, spline smoothing [37]–[39] and polynomial regression [40] as *nonparametric models*, and longitudinal regression [41], generalized linear, and Cox proportional hazard models [42]–[44] as *semiparametric models*.

For ACD, an adequate model assumption and specification is crucial and has many theoretical and applied implications. The main problem is to select a flexible model that can capture nonlinear relations while also providing high detection power and computational efficiency. All these are relevant aspects to consider for the diagnostic measure, for which many methods have been proposed. However, Cook's distance models only linear relations, which limit its applicability to complex real-world data. In recent years, kernel methods have been widely adopted as an appropriate framework for nonlinear model development in machine learning for classification, regression, hypothesis testing, and dimensionality reduction [45], [46]. Kernel methods allow one to derive flexible nonlinear and nonparametric models, are intrinsically regularized, and are endorsed with solid mathematical properties. This has allowed us to define diagnostics based on leveraging the kernel ridge regression (KRR) method [47]. However, despite the excellent modeling performance of KRR, the direct definition of leverage scores based on KRR implies a huge computational cost and the lack of a practical out-of-sample estimates [48]–[50]. This hampers its adoption and usefulness in real practice.

In this article, we introduce the Cook's distance for the KRR model in a reproducing kernel Hilbert space (RKHS) for ACD. Noting the high computational cost, we introduce random Fourier features [51] and the Nyström method [52], [53] for improved efficiency. Both approaches allow us to compute residuals [54] and leverage the KRR *explicitly* in RKHS, while the Nyström method also provides implicit regularization capabilities. Essentially, the Nyström method approximates the large kernel matrix by a much smaller low-rank matrix. Although the best low-rank approximation is obtained by singular value decomposition (SVD), it is computationally expensive. On the contrary, the Nyström method achieves low-rank approximation with considerably higher computationally efficiency [55], [56]. The proposed methods are simple, computationally very efficient in both memory and processing costs, and achieve improved detection compared to standard approaches. We show results in a set of real ACD problems with pairs of large-scale multispectral satellite images acquired by different sensors (Quickbird, Sentinel-2) and involving different changes of interest (floods, wildfires, urbanization, and droughts).

The remainder of this article is organized as follows. Section II sets the notation, introduces the Cook's distance, briefly reviews the concept of influential points and leveraging in statistics, and introduces the direct kernel Cook's distance. Section III elaborates further on our proposed fast implementations and provides a comparison of space and time complexity in all methods. Section IV presents the performance of the proposed fast Cook's chronochrome method for ACD on synthetic and real-world data. Finally, we conclude in Section V with some remarks and prospective future work.

II. Kernelized Cook's Distance

A. Notation and the Chronochrome Approach

Let us define two consecutive d-band multispectral images in matrix form $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^{n\times d}$ composed of n pixels $\mathbf{x}_i, \mathbf{y}_i \in \mathbb{R}^d$, $i = 1, \ldots, n$. Assume that a set of changes have occurred in between, and that such changes do not alter the image distribution significantly. The "chronochrome" approach [30] builds on this idea and fits a model to predict the second image $\mathbf{Y}$ from the first one $\mathbf{X}$ and decides that a point is anomalous (i.e., it has changed) if, for instance, the corresponding residual is significantly large. The prediction function $f : \mathbf{x} \to \mathbf{y}$ is learned from the observations. The task is now to assess the significance of the obtained residuals, $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}$, that is to derive a sensible diagnostic measure.

B. Cook's Distance

Cook's distance comes from the definition of *leverage*, which measures how distant are the independent variable values (of a particular observation) from those of the other observations. The highest leveraged points are those observations that could be considered as extreme or outlying values of the independent variables. Cook's distance measures the effect of removing a given observation. Therefore, the aim is to find out which elements from the sample set are more relevant to the model.

The standard Cook's distance assumes a linear model for prediction of the second image from the first one, i.e., $\hat{\mathbf{Y}} = \tilde{\mathbf{X}}\mathbf{W}$, where $\mathbf{W} \in \mathbb{R}^{(d+1)\times d}$, and $\tilde{\mathbf{X}}$ is the augmented design matrix with a column of ones to account for the bias term, $\tilde{\mathbf{X}} = [\mathbf{X}|\mathbf{1}_n]$. The solution to this least squares problem is given by the Wiener–Hopf normal equations, $\mathbf{W} = (\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}^\top\mathbf{Y}$. The predictions can be expressed as $\hat{\mathbf{Y}} = \tilde{\mathbf{X}}\mathbf{W} = \tilde{\mathbf{X}}(\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}^\top\mathbf{Y} = \mathbf{H}\mathbf{Y}$, where $\mathbf{H}$ is known as the *projection matrix*, and we define the *leverage score* of the ith observation as

$$h_i = \mathbf{x}_i^\top(\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}})^{-1}\mathbf{x}_i. \tag{1}$$

Similarly, the ith element of the residual vector $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = (\mathbf{I} - \mathbf{H})\mathbf{y}$ is denoted as e_i. The Cook's distance D_i for observation $\mathbf{x}_i$, $i = 1, \ldots, n$, is defined as the sum of all the changes in the regression model when the ith observation is deleted

$$D_i = \frac{\sum_{j=1}^{n}\left(\hat{\mathbf{y}}_j - \hat{\mathbf{y}}_{j\backslash i}\right)^2}{d\ \text{MSE}^2} \tag{2}$$

where $\hat{\mathbf{y}}_j$ means to predict the jth sample through the model trained with all the samples and $\hat{\mathbf{y}}_{j\backslash i}$ is the fitted response value obtained when i is excluded, and MSE is the mean square error of the regression model with all samples, i.e., $\text{MSE} = \frac{1}{N}\sum_{j=1}^{n}(\hat{\mathbf{y}}_j - \mathbf{y}_j)^2$. Cook's distance can be equivalently expressed using the leverage

$$D_i = \frac{e_i^2 h_i}{d\ \text{MSE}^2(1-h_i)^2}. \tag{3}$$

Cook showed that this estimation can be obtained using incremental rank-1 updates of covariances, without even needing to recompute each model when the ith sample is removed [32].

C. Kernel Cook's Distance

The kernel Cook's distance can be easily derived by departing from (3). For that, we need to compute both the errors and the leverage scores as a function of the input data only. Let us first recall the KRR prediction formula, $\mathbf{y} = \mathbf{K}(\mathbf{K} + \lambda\mathbf{I})^{-1}\mathbf{y}$, where λ is a regularization parameter, and $\mathbf{K}$ is the kernel matrix. The residuals are thus $\mathbf{e} = (\mathbf{I} - \mathbf{H}^{\mathcal{H}})\mathbf{y}$, where the (kernel) projection matrix $\mathbf{H}^{\mathcal{H}} = \mathbf{K}(\mathbf{K} + \lambda\mathbf{I})^{-1}$, and the (kernel) leveraging scores become

$$h_i^{\mathcal{H}} = \text{diag}(\mathbf{H}^{\mathcal{H}}), \quad i = 1, \ldots, n. \tag{4}$$

From here, one can readily compute $e_i^{\mathcal{H}}$ and the kernel Cook's distance as

$$D_i^{\mathcal{H}} = \frac{(e_i^{\mathcal{H}})^2}{d\ \text{MSE}^2}\frac{h_i^{\mathcal{H}}}{(1-h_i^{\mathcal{H}})^2}. \tag{5}$$

Note that the inversion of a large $\mathbf{K}$ matrix in $\mathbf{H}^{\mathcal{H}}$ has a cost of cubic time complexity and quadratic space (memory) complexity. One could think of computing the leverage scores using an SVD, but the exact computation is as costly as solving the original problem since the cost is also cubic. Unlike the linear case, the recursive solution of (5) is cumbersome, and one has to recompute each model after sample deletion, thus involving a cascade of costly inverse operations.

III. Efficiency in Kernel Cook

In this article, we will exploit both random Fourier features and Nyström approximation of the leverage scores and the errors for Cook's distance approximation.

A. Randomized Cook's Distance

Let us first approximate the kernel matrix with random Fourier features [51]. Formally, we now use a linear regression model expressed on data explicitly projected onto q random Fourier features. Let us define a feature map $\mathbf{z}(\mathbf{x}) : \mathbb{R}^d \to \mathbb{C}^q$, *explicitly* constructed as $\mathbf{z}(\mathbf{x}) := [\exp(\mathbb{i}\mathbf{w}_1^\top\mathbf{x}), \ldots, \exp(\mathbb{i}\mathbf{w}_q^\top\mathbf{x})]^\top$, where $\mathbb{i} = \sqrt{-1}$, and $\mathbf{w}_q \in \mathbb{R}^d$ is randomly sampled from a data-independent distribution [51]. The prediction model is now defined as $\hat{\mathbf{Y}} = \Re\{\mathbf{Z}\mathbf{W}\}$, where $\mathbf{Z} = [\mathbf{z}_1 \cdots \mathbf{z}_n]^\top \in \mathbb{R}^{n\times q}$, with the weight matrix $\mathbf{W} \in \mathbb{R}^{q\times d}$. The *randomized leverage* of a particular sample is now expressed

$$h_i^R = \Re\{\mathbf{z}(\mathbf{x}_i)^\top(\mathbf{Z}^\top\mathbf{Z} + \lambda\mathbf{I})^{-1}\mathbf{z}(\mathbf{x}_i)\} \tag{6}$$

which is then plugged into (3) owing to the linearity of the model where $\mathbf{e}^R = (\mathbf{I} - \mathbf{H}^R)\mathbf{y}$ and then leads to

$$D_i^R = \frac{(e_i^R)^2}{d\ \text{MSE}^2}\frac{h_i^R}{(1-h_i^R)^2}. \tag{7}$$

This allows us to control the memory and computational complexity explicitly through q, as one has to store matrices of $n \times q$ and invert matrices of size $q \times q$ only. It is worth noting that, in practice, a low number of random Fourier features are needed, $q \ll n$. This is not only beneficial in computation time and memory savings but also has a regularization effect in the solution.

B. Nyström Cook's Distance

The Nyström method selects a small set of $r \ll n$ samples to make a low-rank approximation of an $n \times n$ kernel matrix $\mathbf{K} \approx \mathbf{K}_{rn}^\top\mathbf{K}_{rr}^{-1}\mathbf{K}_{rn}$ [52], where $\mathbf{K}_{rn} \in \mathbb{R}^{r\times n}$ contains the kernel similarities between $\hat{\mathbf{X}} \in \mathbb{R}^{r\times d}$ and $\mathbf{X} \in \mathbb{R}^{n\times d}$, and $\mathbf{K}_{rr} \in \mathbb{R}^{r\times r}$ is a kernel matrix containing data similarities between the points in $\hat{\mathbf{X}}$. By exploiting the Nyström method in the Woodbury–Morrison formula, we obtain

$$(\mathbf{K} + \lambda\mathbf{I})^{-1} = \lambda^{-1}(\mathbf{I} - \mathbf{K}_{nr}(\lambda\mathbf{K}_{rr} + \mathbf{K}_{nr}^\top\mathbf{K}_{nr})^{-1}\mathbf{K}_{nr}^\top) \tag{8}$$

and now defining $\mathbf{Q} = \lambda\mathbf{K}_{rr} + \mathbf{K}_{nr}^\top\mathbf{K}_{nr}$, the projection matrix approximation is defined as

$$\mathbf{H}^N = \lambda^{-1}\mathbf{K}(\mathbf{I} - \mathbf{K}_{nr}\mathbf{Q}^{-1}\mathbf{K}_{nr}^\top) \tag{9}$$

TABLE I
SPACE AND TIME COMPLEXITY FOR ALL METHODS

Method	**T**	**C**	$\mathbf{C^{-1}}$	**W**	**L**	**ACD**	$\mathcal{O}(.)$
Space							
L-Cook	–	d^2	d^2	d^2	n	n	$\mathcal{O}(nd)$
R-Cook	nq	q^2	q^2	q^2	n	n	$\mathcal{O}(nq)$
N-Cook	n^2	r^2	r^2	–	n	n	$\mathcal{O}(n^2)$
K-Cook	n^2	n^2	n^2	–	n	n	$\mathcal{O}(n^2)$
Time							
L-Cook	–	nd^2	d^3	nd^2	nd^2	nd^2	$\mathcal{O}(nd^2)$
R-Cook	nqd	nq^2	q^3	nq^2	nq^2	nq^2	$\mathcal{O}(nq^2)$
N-Cook	n^2d	nr^2	r^3	–	n^2r	n^2d	$\mathcal{O}(n^2r)$
K-Cook	n^2d	n^3	n^3	–	n^3	n^2d	$\mathcal{O}(n^3)$

$\mathbf{T}$ is transformation of image into a nonlinear space, $\mathbf{C}$ is for covariance/kernel matrix, $\mathbf{W}$ is for regression weight, $\mathbf{L}$ is for leverage, $\mathbf{ACD}$ is the Cook's distance, and $\mathcal{O}(.)$ is the overall complexity.

with Nyström leverage scores

$$h_i^N = \text{diag}(\mathbf{H}^n) \tag{10}$$

and $\mathbf{e}^N = (\mathbf{I} - \mathbf{H}^N)\mathbf{y}$; thus, the Nyström Cook's distance becomes

$$D_i^N = \frac{(e_i^N)^2}{d\ \text{MSE}^2} \frac{h_i^N}{(1-h_i^N)^2}. \tag{11}$$

C. Memory and Computational Cost

Space (memory) and time (computational) efficiency of the linear and nonlinear versions are presented in Table I. In this study, the linear version is named as L-Cook, while the nonlinear versions are named Randomized Cook (R-Cook), Nyström Cook (N-Cook), and Kernel Cook (K-Cook). Note that d is the spectral dimension, and it is around 10 for multispectral images and around 100 for hyperspectral images. Although q and r can have similar values, generally $q < r$. Since large images are used, n is much larger than r, q, and d. Therefore, in general, $d < q < r \ll n$.

As can be seen in Table I, the L-Cook method provides superior space and time efficiency. However, the L-Cook method is only limited to rare linear scenarios, where the real-world nonlinear transformation between multitemporal images are formed due to various reasons. However, space and time complexity of the K-Cook method is proportional to the number of pixels in the image, respectively, quadratic in space and cubic in time. Thus, the use of the K-Cook method is not feasible for large images, which is the common scenario nowadays. Note that, for the N-Cook method, kernel matrix $\mathbf{K}$ is still used in (9), but there is no inversion operation on it. Therefore, N-Cook has same space complexity with the K-Cook method as we need to store kernel matrix $\mathbf{K}$. However, time complexity of N-Cook is still superior to the K-Cook method, since only an $r \times r$ matrix is inverted.

IV. EXPERIMENTAL RESULTS

This section analyzes the performance of the proposed linear and nonlinear Cook's distance methods for ACD. In order to test the robustness of the proposed methods, we performed tests in both simulated and real scenes with changes. We evaluate the detection performance of the methods quantitatively through the area under the curve (AUC) of the receiver operating characteristic (ROC) and qualitatively by inspection of the detection maps. We have performed two experiments with different complexities of difficulty while controlling the analyzed changes. The first experiment is designed over a real scenario and synthetic changes. The second set of experiments deals with both real scenes and natural changes related to floods, fires, and urbanization. In order to ease the reproducibility, we provide MATLAB implementations of the methods. Moreover, we made available a database with the labeled images used in the second experiment.[1]

A. Experiment 1: Real Scene With Simulated Changes

The aim of this experiment is to show and analyze the performance of the proposed methods when the change between images is nonlinearly distributed. In this example, we can analyze how nonlinear methods fit the regression model to the data well and how they detect the influential points in the Cook's distance approach. The experiment involves representing a nonlinear relation between two images in order to demonstrate the limitations of the linear algorithms in this situation.

Fig. 1(a) and (b) shows an aerial scene taken over the Image Processing Laboratory from Google Earth in the R band. Fig. 1(a) represents the image at time t_1 (no change class), while Fig. 1(b) represents the image at time t_2 (change class). All the values of the second image (t_2) were modified by applying a soft nonlinear function (an inverted parabola) to simulate nonanomalous changes. In order to introduce the anomalous changes, we interchanged square patches of 4×4 pixels randomly selected.

Since kernel Cook's distance is computationally very demanding, we have selected a portion of the full image in order to have a comparison of all proposed methods together. In particular, we used the region of interest shown in Fig. 1(c) and marked in a red box in Fig. 1(b); the anomalies are highlighted in a black rectangle and the anomalous class represent 0.016%. Fig. 1(d) represents the scatter of original image x-axis against transformed image y-axis; the points in yellow color are the change pixels, but the points in blue color ideally would not be detected as an anomalous change pixels. Fig. 1(e) illustrates how a linear model does not fit the distribution well and the inferred values lead to false-positive errors (in the tails) and true negative errors (green color). Fig. 1(f) shows how a nonlinear model over distribution fits well and both avoid the false positives and detect the changed pixels in the images. These results are confirmed visually through the prediction maps in Fig. 2, where the kernel Cook's distance excels in detection.

In contrast, Table II showcases how efficient the proposed efficient methods can be, achieving better values of AUC compared to the kernel one in less time. Therefore, because of the huge computational cost involved in its calculation, one cannot use it in standard images (even as small as the one in Fig. 1), so efficient algorithms for computing Cook's distances in nonlinear kernel settings are strictly necessary.

[1][Online]. Available: http://isp.uv.es/code/kcook

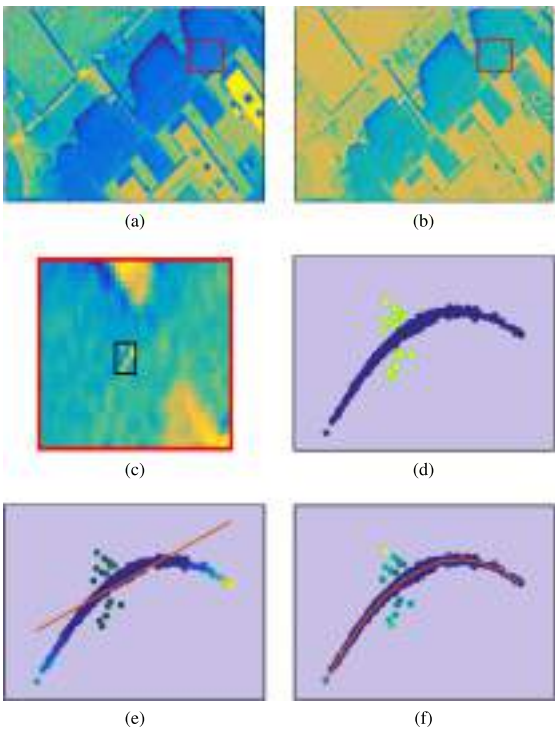


Fig. 1. (a) Image (R band) at time t_1 and the region of interest (red box). (b) Image (R band) at time t_2 and the region of interest (red box). We apply background color distortion and added square patches of 4×4 over t_2 simulating the anomalies, (c) region of interest (red box in t_2) and the corresponding label is surrounded and highlighted in black, (d) scatter plots between t_1 and t_2 pixels in R band, blue dots represent the non-change class and the yellow dots correspond to change class. Panel (e) shows how mis-specification of the linear regression model cannot detect anomalies, while a nonlinear Cook's distance can in (f). In both (e) and (f), the dots color specify how much anomalous the point is for the model (blue less, yellow more).

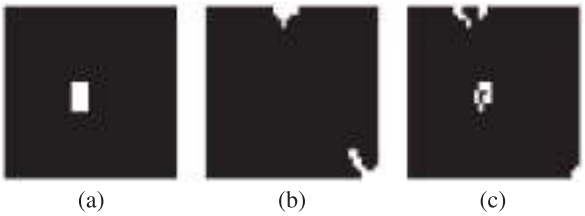


Fig. 2. (a) Prediction map (labels). (b) Change prediction map detected by the linear method. (c) Change prediction map detected by the nonlinear Cook's distance.

B. Experiment 2: Real and Natural Changes

In this section, we report experiments in several real satellite images. We aim to detect changes that can be found naturally in a real environment. The dataset is composed of five different scenes with natural changes, including urbanization, wildfires, droughts, and flooding.

1) Data Collection: We collected pairs of multispectral images acquired at different times over the same location. We selected the images in such a way that a noticeable change happened between the two acquisition times. We photo-interpreted and manually labeled all the image pixels affected by a change of interest. This step is critical and delicate since we could fall into many false alarms due to, for instance, shadows, illumination changes, or natural changes in the vegetation. All images contain changes of a different nature, which allows us to study how the different Cook's distance algorithms perform in a diversity of realistic scenarios.

TABLE II
AUC AND THEIR RESPECTIVE TIME VALUES (IN SECONDS) PER METHOD

Methods	L-Cook	R-Cook	N-Cook	K-Cook
AUC	0.55	0.93	0.93	0.92
Time	0.01	0.03	2.64	6.32

TABLE III
IMAGE ATTRIBUTES USED IN THE EXPERIMENTATION DATASET

Images	**Sensor**	**Size**	**Bands**	**SR**
Argentina	Sentinel-2	381 x 500	12	10-60m
Denver	Quickbird	101 x 101	4	0.6-2.4m
Arizona	Cross-Sensor	201 x 201	7	30m
Texas	Cross-Sensor	301 x 201	7	30m
Australia	Sentinel-2	201 x 501	12	10-60m

A brief summary of the images and change events is as follows. The Argentina dataset represents an area burned between the months of July and August 2016. Denver Region Urbanized Project Area describes the stereo-compiled building roofprints feature of Denver Regional Council of Government. The Texas wildfire dataset is composed of a set of four images acquired by different sensors over Bastrop County, Texas (USA) and is composed of a Landsat 5 TM as the pre-event image and a Landsat 5 TM plus an EO-1 ALI and a Landsat 8 as postevent images. This phenomenon is considered the most destructive wildland–urban interface wildfire in Texas history. The Arizona dataset corresponds to the decline of Lake Powell in the USA. The first image was taken by Landsat-5 and shows its highest water level. The second was taken by Landsat-8 following a period of drought that began in 2000. When the water volume was measured five months later, it was less than half of the maximum lake capacity. The Australia dataset shows the natural floods caused by Cyclone Debbie in Australia 2017. Storm damage resulted from both the high winds associated with the cyclone, and the very heavy rain that produced major riverine floods. Table III gives some descriptors of the images in the database, while Fig. 3 shows the RGB composites of the pairs of images and the corresponding reference map.

2) Numerical Comparison: We selected the hyperparameters using 1000 randomly selected pixels for cross-validation. Each method implies a different set of parameters. For both the randomized and Nyström methods, we have cross-validated the r and q parameters by exploring values between 1 and 400, particularly $r, q \in \{1, 5, 10, 25, 50, 100, 200, 300, 400\}$. In this work, we used the standard radial basis function (RBF) kernel function, $k(\mathbf{x}, \mathbf{x}') = \exp(-\|\mathbf{x} - \mathbf{x}'\|^2/(2\sigma^2))$. The RBF kernel shows good theoretical properties (universal kernel, smoothness, and robustness), convenience (only the lengthscale parameter σ needs to be tuned), and good performance in practice. The RBF

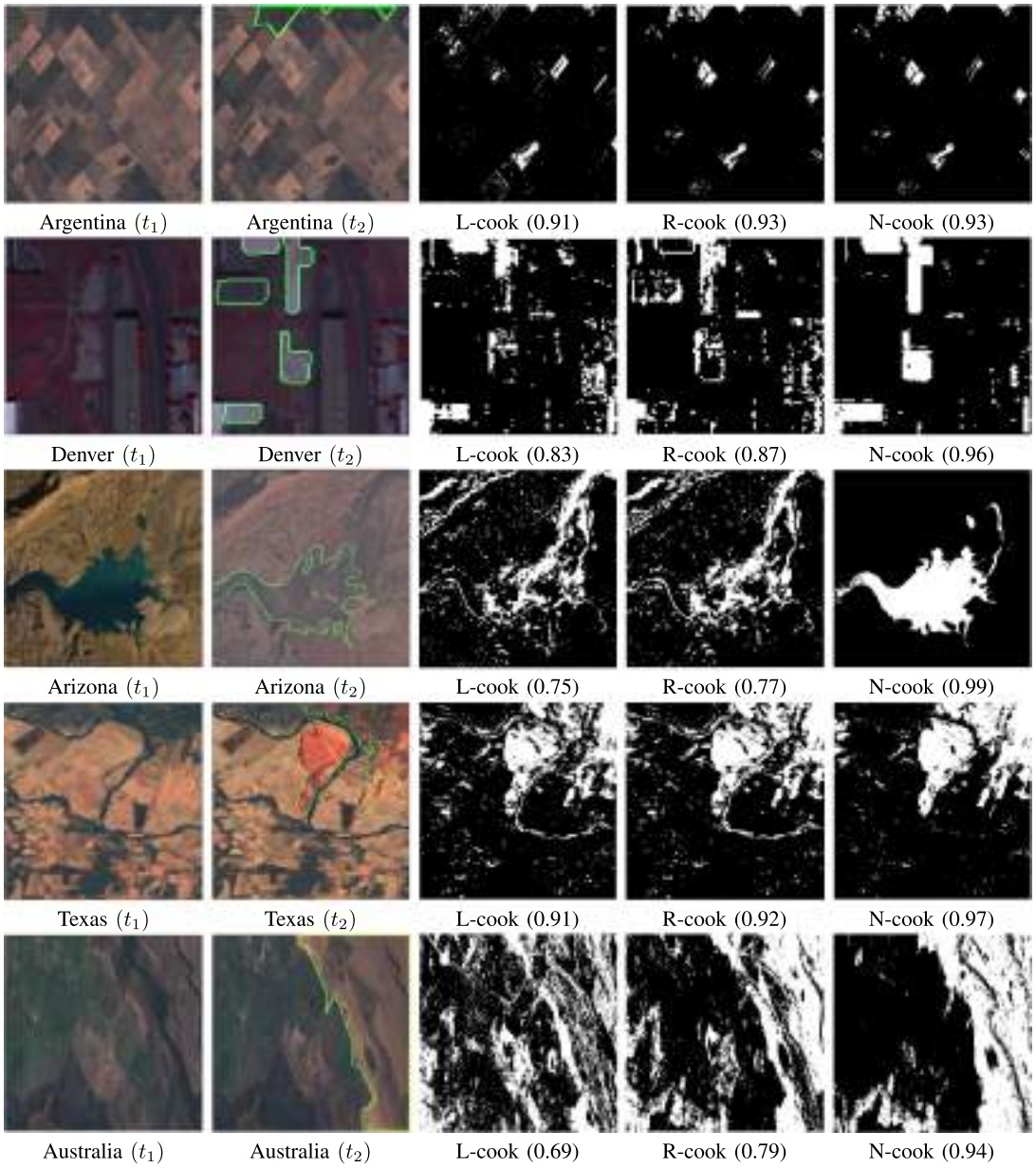


Fig. 3. Images with natural changes and predictions maps. First row: area burned in Argentina between the months of July and August 2016, anomalous samples represent 2.7%. Second row: urbanization area over Denver city correspond to roofprints (extension of anomalous pixels represents the 11.5% of the image). Third row: decline of the Lake Powell in Arizona, USA (16.35%). Fourth row: the most destructive wildland-urban interface wildfire in Texas history (19.5%). Last row: natural floods caused by Cyclone Debbie in Australia (34%). First column: images without changes, first time of acquisition (t_1). Second column: images with the anomalous changes and their corresponding labels are surrounded and highlighted with green color, second time of acquisition (t_2). Third column: prediction map of linear method. Fourth column: prediction map of random Fourier features method. Last column: prediction map of Nyström approximation method. AUC value in parentheses.

kernel is used to perform kernel regression, which incorporates a regularization parameter λ. We searched both σ and λ logarithmic grids between 10^{-4} and 10^{20}.

We optimized the hyperparameters of different methods to maximize the cross-validation AUC. We compared the ROCs and precision–recall curves in terms of AUCs for all methods and images in Fig. 4 . In general, all methods can cope with the large dimension of the images and can provide reasonable results, AUC> 0.70 (see Table IV).

The nonlinear versions (randomized and Nyström approximations) improve the results of the linear Cook's distance, revealing nonlinear changes in all scenes, yet differences are minor for the Texas scene. The Nyström Cook's distance achieves consistently the best results in all the scenarios, and false or positive rates

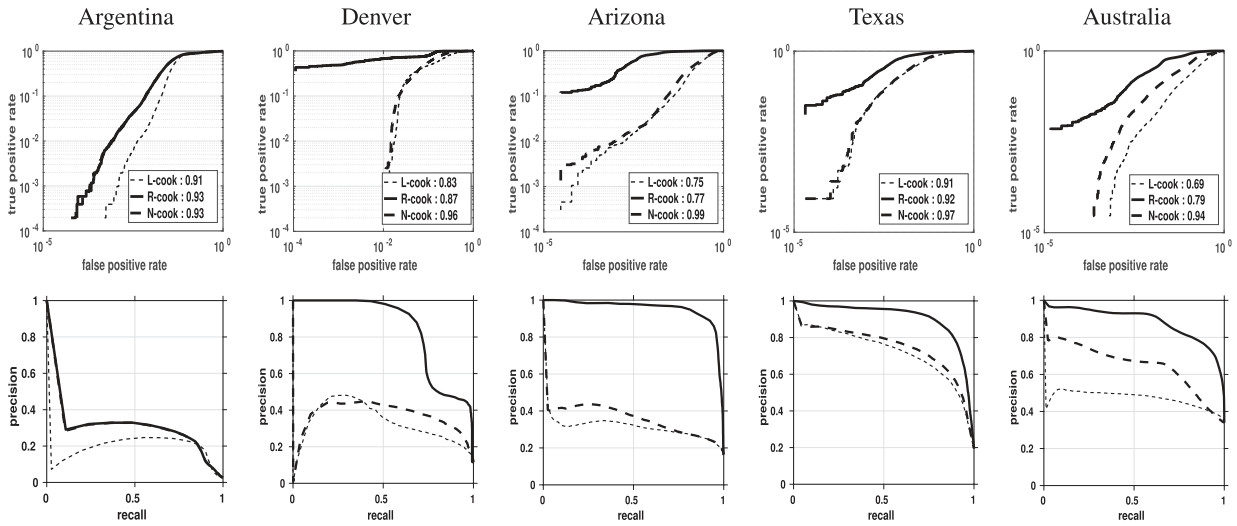


Fig. 4. ROC curves and precision–recall for all images by columns. First row showcases the ROC curves in logarithmic scale. Numbers in legend display the AUC values for each method. Second row showcases the precision–recall following the ROC curves legend.

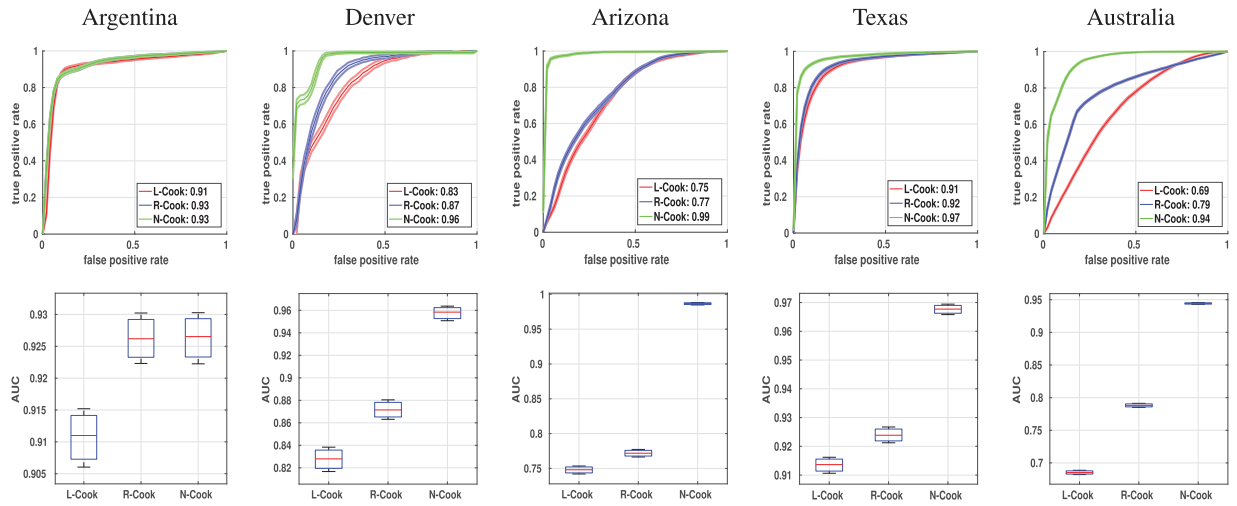


Fig. 5. Bootstrap experiment. In the top row, mean value of the 1000 experiments is plotted as ROC curves with the standard deviation of each detection algorithm represented by the shaded region. In the bottom row, AUC values and standard deviation for each method are shown as boxplot.

TABLE IV
AUC PER METHOD AND SCENE

Methods	Argentina	Denver	Arizona	Texas	Australia
L-Cook	0.91	0.83	0.75	0.91	0.69
R-Cook	0.93	0.87	0.77	0.92	0.79
N-Cook	**0.93**	**0.96**	**0.99**	**0.97**	**0.94**

The best results are bold faced.

regimes. A average gain of +15.6% over the linear approach and of +11.8% over the randomized approach, along with the computational efficiency, justify the adoption of this approach. The double logarithmic plot aims to better appreciate the differences in very low false positive rate regimes. In addition, precision and recall are an understanding and measure of relevance. Here, it becomes clear that the Nyström approach excels in all images.

For each experiment, 1000 runs were made for testing the significance of the methods based on the ROC profiles. The mean value of the experimental runs is plotted with the standard deviation of each detection algorithm represented by the shaded region in Fig. 5. In addition, a boxplot is shown in the same figure to illustrate the standard deviation of each methods with a better precision. As seen in Fig. 5, N-Cook has always superior or equivalent performance compared to L-Cook and R-Cook, i.e., higher detection rate and lower false alarm rate, and higher AUC value and lower standard deviation.

3) Visual Comparison: A visual comparison of the results is given in Fig. 3. Differences between the L-Cook and the

R-Cook are not visually significant either. In general, N-Cook yields clear and sharper detection maps (last column), especially in large spatial structures (see, e.g., roofs in Denver, lake in Arizona) but also exhibits a much lower false alarm rate (see, e.g., a less amount of spurious detections in Texas wildfires). This is, however, sometimes compensated with sensitivity to subtle reflectance changes and misclassified pixels in Australia due to imperfect labeling of pixels. This is why this problem is so difficult to solve in an automatic way.

V. Conclusion

We introduced the kernel Cook's distance for ACD settings, with the particular focus on RS image CD problems. The key in the proposed methodology is to redefine the ACD problem in an RKHS where the data are mapped to. This endorses the methods with improved capacity and flexibility, since nonlinear feature relations (and hence outliers) can be better identified. However, the obtained kernelized method encounters huge computational problems in practice, which hampers its applicability and wider adoption. To resolve this problem, we proposed computationally efficient techniques based on random Fourier features and low-rank Nyström approximations and compared their capabilities in a wide range of both simulated and real changes. The Nyström approximation excelled over the rest of the implementations in both simulated and real scenarios and in terms of accuracy and efficiency. Future work will study other related kernel diagnostic measures. Extension to online and multichange problems are also topics of further research.

References

[1] S. Liu, D. Marinelli, L. Bruzzone, and F. Bovolo, "A review of change detection in multitemporal hyperspectral images: Current techniques, applications, and challenges," *IEEE Geosci. Remote Sens. Mag.*, vol. 7, no. 2, pp. 140–158, Jun. 2019.

[2] A. Asokan and J. Anitha, "Change detection techniques for remote sensing applications: A survey," *Earth Sci. Inf.*, vol. 12, pp. 1–18, 2019.

[3] S. Pouyanfar *et al.*, "Unconstrained flood event detection using adversarial data augmentation," in *Proc. IEEE Int. Conf. Image Process.*, 2019, pp. 155–159.

[4] P. Mishra, T. Feller, M. Schmuck, A. Nicol, and A. Nordon, "Early detection of drought stress in *Arabidopsis thaliana* utilsing a portable hyperspectral imaging setup," in *Proc. Workshop Hyperspectral Imag. Signal Process., Evol. Remote Sens.*, 2019, pp. 1–5.

[5] P. Du, S. Liu, P. Gamba, K. Tan, and J. Xia, "Fusion of difference images for change detection over urban areas," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 5, no. 4, pp. 1076–1086, Aug. 2012.

[6] G. Camps-Valls, D. Tuia, L. Bruzzone, and J. A. Benediktsson, "Advances in hyperspectral image classification: Earth monitoring with statistical learning methods," *IEEE Signal Process. Mag.*, vol. 31, no. 1, pp. 45–54, Jan. 2014.

[7] A. Plaza *et al.*, "Recent advances in techniques for hyperspectral image processing," *Remote Sens. Environ.*, vol. 113, no. Suppl. 1, pp. S110–S122, 2009.

[8] G. Camps-Valls, D. Tuia, L. Gmez-Chova, S. Jimnez, and J. Malo, *Remote Sensing Image Processing*, 1st ed. San Rafael, CA, USA: Morgan & Claypool, 2011.

[9] J. Theiler and S. Perkins, "Proposed framework for anomalous change detection," in *Proc. ICML Workshop Mach. Learn. Algorithms Surveillance Event Detection*, Jan. 2006, pp. 7–14.

[10] J. Theiler, "Quantitative comparison of quadratic covariance-based anomalous change detectors," *Appl. Opt.*, vol. 47, no. 28, pp. F12–F26, 2008.

[11] I. Steinwart, J. Theiler, and D. Llamocca, "Using support vector machines for anomalous change detection," *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2010, pp. 3732–3735.

[12] J. Theiler, C. Scovel, B. Wohlberg, and B. R. Foy, "Elliptically contoured distributions for anomalous change detection in hyperspectral imagery," *IEEE Geosci. Remote Sens. Lett.*, vol. 7, no. 2, pp. 271–275, Apr. 2010.

[13] J. Theiler and A. M. Matsekh, "Total least squares for anomalous change detection," *Proc. SPIE*, vol. 7695, pp. 507–518, 2010.

[14] J. Theiler and B. Wohlberg, "Local coregistration adjustment for anomalous change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 50, no. 8, pp. 3107–3116, Aug. 2012.

[15] J. Theiler, "Spatio-spectral anomalous change detection in hyperspectral imagery," in *Proc IEEE Global Conf. Signal Inf. Process.*, 2013, pp. 953–956.

[16] J. Theiler and A. Ziemann, "Background estimation in multispectral imagery," in *Proc. Opt. Sens. Sens. Congr.*, 2019, Paper HW6B.1.

[17] J. Simmons, L. Drummy, C. Bouman, and M. Graef, *Statistical Methods for Materials Science: The Data Science of Microstructure Characterization*. Boca Raton, FL, USA: CRC Press, 2019.

[18] C. Wu, B. Du, and L. Zhang, "Hyperspectral anomalous change detection based on joint sparse representation," *ISPRS J. Photogrammetry Remote Sens.*, vol. 146, pp. 137–150, 2018.

[19] M. A. F. Pimentel, D. A. Clifton, L. Clifton, and L. Tarassenko, "A review of novelty detection," *Signal Process.*, vol. 99, pp. 215–249, 2014.

[20] N. Longbotham and G. Camps-Valls, "A family of kernel anomaly change detectors," in *Proc. 6th Workshop Hyperspectral Image Signal Process.: Evol. Remote Sens.*, 2014, pp. 1–4.

[21] Chris Clifton, "Change detection in overhead imagery using neural networks," *Appl. Intell.*, vol. 18, pp. 215–234, 2003.

[22] J. A. P. Hidalgo, A. Pérez-Suay, F. Nar, and G. Camps-Valls, "Nonlinear Cook distance for anomalous change detection," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2018, pp. 5025–5028.

[23] J. A. P. Hidalgo, V. Laparra, N. Longbotham, and G. Camps-Valls, "Kernel anomalous change detection for remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 10, pp. 7743–7755, Oct. 2019.

[24] Y. Elhadad, S. R. Rotman, and D. Blumberg, "Analysis of hyperspectral anomaly change detection algorithms," in *Proc. Workshop Hyperspectral Image Signal Process., Evol. Remote Sens.*, 2016, pp. 1–5.

[25] N. Acito, M. Diani, G. Corsini, and S. Resta, "Introductory view of anomalous change detection in hyperspectral images within a theoretical Gaussian framework," *IEEE Aerosp. Electron. Syst. Mag.*, vol. 32, no. 7, pp. 2–27, Jul. 2017.

[26] R. Mayer, F. Bucholtz, and D. Scribner, "Object detection by using 'whitening/dewhitening' to transform target signatures in multitemporal hyperspectral and multispectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 41, no. 5, pp. 1136–1142, May 2003.

[27] J. Arenas-Garcia, K. B. Petersen, G. Camps-Valls, and L. K. Hansen, "Kernel multivariate analysis framework for supervised subspace learning: A tutorial on linear and kernel multivariate methods," *IEEE Signal Process. Mag.*, vol. 30, no. 4, pp. 16–29, Jul. 2013.

[28] A. A. Green, M. Berman, P. Switzer, and M. D. Craig, "A transformation for ordering multispectral data in terms of image quality with implications for noise removal," *IEEE Trans. Geosci. Remote Sens.*, vol. 26, no. 1, pp. 65–74, Jan. 1988.

[29] A. A. Nielsen, K. Conradsen, and J. J. Simpson, "Multivariate alteration detection MAD and MAF postprocessing in multispectral, bitemporal image data: New approaches to change detection studies," *Remote Sens. Environ.*, vol. 64, no. 1, pp. 1–19, 1998.

[30] A. Schaum and A. Stocker, "Long-interval chronochrome target detection," in *Proc. Int. Symp. Spectral Sensing Res.*, 1997.

[31] A. Schaum and A. Stocker, "Linear chromodynamics models for hyperspectral target detection," in *Proc. IEEE Aerosp. Conf. Proc.*, 2003, vol. 4, pp. 1879–1885.

[32] R. D. Cook, "Detection of influential observation in linear regression," *Technometrics*, vol. 19, no. 1, pp. 15–18, 1977.

[33] R. D. Snee, "Regression diagnostics: Identifying influential data and sources of collinearity," *J. Qual. Technol.*, vol. 15, no. 3, pp. 149–153, 1983.

[34] R. D. Cook and S. Weisberg, "Characterizations of an empirical influence function for detecting influential cases in regression," *Technometrics*, vol. 22, no. 4, pp. 495–508, 1980.

[35] A. E. Hoerl and R. W. Kennard, "Ridge regression: Biased estimation for nonorthogonal problems," *Technometrics*, vol. 12, no. 1, pp. 55–67, 1970.

[36] Tibshirani R, "Regression shrinkage and selection via the Lasso," *J. Royal Statist. Soc., Ser. B (Methodol.)*, vol. 58, no. 1, pp. 267–288, 1996.

[37] R. L. Eubank, "Diagnostics for smoothing splines," *J. Royal Statist. Soc. Ser. B (Methodol.)*, vol. 47, no. 2, pp. 332–341, 1985.
[38] K. Choongrak and E. S. Barry, "Reference values for Cook's distance," *Commun. Statist.—Simul. Comput.*, vol. 25, no. 3, pp. 691–708, 1996.
[39] B. W. Silverman, *Density Estimation for Statistics and Data Analysis*. London, U.K.: Chapman & Hall, 1986.
[40] C. Kim, Y. Lee, and B. U. Park, "Cook's distance in local polynomial regression," *Statist.& Probab. Lett.*, vol. 54, no. 1, pp. 33–40, 2001.
[41] W. Bae, S. Hwang, and C. Kim, "Influence diagnostics in the varying coefficient model with longitudinal data," *Comput. Statist.*, vol. 23, pp. 185–196, 2008.
[42] H. Zhu, J. G. Ibrahim, and M.-H. Chen, "Diagnostic measures for the Cox regression model with missing covariates," *Biometrika*, vol. 102, no. 4, pp. 907–923, 2015.
[43] H. Zhu, J. G. Ibrahim, and X. Shi, "Diagnostic measures for generalized linear models with missing covariates," *Scand. J. Statist.*, vol. 36, no. 4, pp. 686–712, 2009.
[44] H. Zhu, S. Lee, B. Wei, and J. Zhou, "Case-deletion measures for models with incomplete data," *Biometrika*, vol. 88, no. 3, pp. 727–737, 2001.
[45] B. Schölkopf and A. Smola, *Learning with Kernels—Support Vector Machines, Regularization, Optimization and Beyond*. Cambridge, MA, USA: MIT Press, 2002.
[46] J. L. Rojo-Álvarez, M. Martínez-Ramón, J. Munoz Marí, and G. Camps-Valls, *Digital Signal Processing with Kernel Methods*. Hoboken, NJ, USA: Wiley, 2017.
[47] J. Shawe-Taylor and N. Cristianini, *Kernel Methods for Pattern Analysis*. Cambridge, MA, USA: Cambridge Univ. Press, 2004.
[48] A. Alaoui and M. W. Mahoney, "Fast randomized kernel ridge regression with statistical guarantees," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, 2015, pp. 775–783.
[49] Alessandro Rudi, Daniele Calandriello, Luigi Carratino, and Lorenzo Rosasco, "On fast leverage score sampling and optimal learning," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, 2018, pp. 5672–5682.
[50] S. McCurdy, "Ridge regression and provable deterministic ridge leverage score sampling," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, 2018, pp. 2463–2472.
[51] A. Rahimi and B. Recht, "Random features for large-scale kernel machines," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, 2007, pp. 1177–1184.
[52] C. K. I. Williams and M. Seeger, "Using the Nyström method to speed up kernel machines," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, 2000, pp. 682–688.
[53] C. Zhao, G. Zhao, and X. Jia, "Hyperspectral image unmixing based on fast kernel archetypal analysis," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 1, pp. 331–346, Jan. 2017.
[54] R. Touati, M. Mignotte, and M. Dahmane, "Anomaly feature learning for unsupervised change detection in heterogeneous images: A deep sparse residual model," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 588–600, 2020.
[55] T. Yang, Y.-F. Li, M. Mahdavi, R. Jin, and Z.-H. Zhou, "Nyström method vs random fourier features: A theoretical and empirical comparison," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, 2012, pp. 476–484.
[56] N. K. Kumar and J. Schneider, "Literature survey on low rank approximation of matrices," *Linear Multilinear Algebra*, vol. 65, no. 11, pp. 2212–2244, 2017.

José Antonio Padrón-Hidalgo received the B.Sc. degree in telecommunications and electronics from the University of Pinar del Río, Pinar del Río, Cuba, in 2015. He is currently working toward the Ph.D. degree in electronics with the Universitat de València, València, Spain.

His current research interests include developing algorithms in order to detect anomalous and extreme changes for remote sensing imagery.

Adrián Pérez-Suay received the B.Sc. degree in mathematics, the master's degree in advanced computing and intelligent systems, and the Ph.D. degree in computational mathematics and computer science from the Universitat de València, València, Spain, in 2007, 2010, and 2015, respectively.

He is currently an Assistant Professor with the Department of Mathematics Education, Universitat de València, where he is also a Researcher with the Image and Signal Processing Group. His research interests include dependence estimation, kernel methods, and causal inference for remote sensing data analysis.

Fatih Nar received the Ph.D. degree in information systems from the Middle East Technical University, Ankara, Turkey.

From 1996 to 2016, he was a Software Developer, Database Administrator, Trainer, Technical Leader, and Researcher with many software development and research companies. From 2017 to 2018, he was a visiting Postdoctoral Researcher with the Image Processing Laboratory, Universitat de València, València, Spain. From 2016 to 2020, he was an Assistant Professor with the Department of Computer Engineering, Konya Food and Agriculture University, Konya, Turkey. In 2020, he joined the Department of Computer Engineering, Ankara Yıldırım Beyazıt University, Ankara, as an Assistant Professor. His main research interests include image processing, machine learning, and remote sensing.

Valero Laparra was born in València, Spain, in 1983. He received the B.Sc. degree in telecommunications engineering and the B.Sc. degree in the electronics engineering from the Universitat de València, València, in 2005 and 2007, respectively, the B.Sc. degree in mathematics from the Universidad Nacional de Educación a Distancia, Madrid, Spain, in 2010, and the Ph.D. degree in computer science and mathematics from the Universitat de València, in 2011.

He is currently an Assistant Professor with the Escuela Técnica Superior de Ingenería, Universitat de València, where he is also a Researcher with the Image Processing Laboratory.

Gustau Camps-Valls received the Ph.D. degree in physics from the Universitat de València, València, Spain, in 2002.

He is currently a Full Professor of Electrical Engineering and a Coordinator of the Image and Signal Processing Group, Universitat de València. He is involved in the development of machine learning algorithms for geoscience and remote sensing data analysis. He has authored 200 journal papers, more than 200 conference papers, and 20 international book chapters. He holds a Hirsch's index, h = 60 (source: Google Scholar), entered the ISI list of Highly Cited Researchers, in 2011, and Thomson Reuters ScienceWatch identified one of his papers on kernel-based analysis of hyperspectral images as a fast moving front research. He is the Editor for the books entitled *Kernel Methods in Bioengineering, Signal and Image Processing* (Hershey, PA, USA: IGI, 2007), *Kernel Methods for Remote Sensing Data Analysis* (Hoboken, NJ, USA: Wiley, 2009), *Remote Sensing Image Processing* (San Rafael, CA, USA: Morgan & Claypool, 2011), and *Digital Signal Processing with Kernel Methods* (Hoboken, NJ, USA: Wiley, 2018).

Dr. Camps-Valls was a recipient of the Prestigious European Research Council Consolidator Grant on Statistical Learning for Earth Observation Data Analysis, in 2015. He is/has been an Associate Editor for the IEEE TRANSACTIONS ON SIGNAL PROCESSING, IEEE GEOSCIENCE AND REMOTE SENSING LETTERS, and IEEE SIGNAL PROCESSING LETTERS. He was the Invited Guest Editor for the IEEE JOURNAL OF SELECTED TOPICS IN SIGNAL PROCESSING, in 2012 and *IEEE Geoscience and Remote Sensing Magazine*, in 2015.

Kernel Anomalous Change Detection for Remote Sensing Imagery

José A. Padrón-Hidalgo, Valero Laparra, Nathan Longbotham, and Gustau Camps-Valls, *Fellow, IEEE*

***Abstract*— Anomalous change detection (ACD) is an important problem in remote sensing image processing. Detecting not only pervasive but also anomalous or extreme changes has many applications for which methodologies are available. This paper introduces a nonlinear extension of a full family of anomalous change detectors. In particular, we focus on algorithms that utilize Gaussian and elliptically contoured (EC) distribution and extend them to their nonlinear counterparts based on the theory of reproducing kernels' Hilbert space. We illustrate the performance of the kernel methods introduced in both pervasive and ACD problems with real and simulated changes in multispectral and hyperspectral imagery with different resolutions (AVIRIS, Sentinel-2, WorldView-2, and Quickbird). A wide range of situations is studied in real examples, including droughts, wildfires, and urbanization. Excellent performance in terms of detection accuracy compared to linear formulations is achieved, resulting in improved detection accuracy and reduced false-alarm rates. Results also reveal that the EC assumption may be still valid in Hilbert spaces. We provide an implementation of the algorithms as well as a database of natural anomalous changes in real scenarios http://isp.uv.es/kacd.html.**

***Index Terms*— Anomalous change detection (ACD), elliptical distributions, Gaussianity, hyperbolic ACD, kernel methods.**

I. Introduction

THE problem of change detection deals with identifying transitions between a pair (or a series) of coregistered images [1], [2]. Change detection in remote sensing images is of paramount relevance because it automates traditional manual tasks in disaster management (floods, droughts, and wildfires) and it helps in designing development and settlement plans as well as in urban and crop monitoring. Multitemporal classification and change detection are very active fields nowadays because of the increasingly available complete time series of images and the interest in monitoring changes occurring on the earth's cover due to either natural or anthropogenic activities. Complete constellations of civil and military satellite sensors currently provide high spatial resolution and high revisiting frequency. The Copernicus' Sentinels[1] or NASA's A-train[2] programs are producing near real-time coverage of the globe. NASA is currently producing a Harmonized Landsat Sentinel-2 (HLS) data set, which can be used for monitoring agricultural resources with an unprecedented combination of 30-m spatial resolution and two to three days revisit. In parallel, new commercial satellite missions are being deployed to provide multispectral data at both high-spatial and high-temporal resolutions. For example, the PlanetScope constellation by Planet Labs, Inc., can provide 5-m data daily for sites requested by the client, and the recently announced UrtheDaily constellation, specifically designed for operational agricultural applications, will acquire S2-like data also at 5-m spatial resolution and with full global coverage every day. It goes without saying that closed-range applications using drones and all kind of unmanned automated vehicles (UAVs) also challenge the field of automatic change detection. All in all, automatic image analysis in general and change detection, in particular, are becoming necessary in the current era of data deluge.

An interesting and related problem is that of anomalous change detection (ACD): this configuration differs from standard change detection in that the objective is to identify only rare (or anomalous) events, ignoring the regular ones. In this paper, we extend the family of ACD methods in [3] to cope with higher order feature relations through the theory of reproducing kernels. Kernel methods allow the generalization of algorithms that are expressed in terms of dot products to account for higher order (nonlinear) feature relations, yet still relying on linear algebra [4]–[7]. We illustrate the performance of the introduced kernel ACD methods in different experiments involving synthetic, artificially enforced, and natural anomalous changes in multispectral and hyperspectral imagery with different spatial resolutions (AVIRIS, Sentinel-2, WorldView-2, and Quickbird). A wide range of situations is studied, involving droughts, wildfires, and urbanization in real examples. Very good performance is achieved in terms of detection accuracy compared to the linear formulations. Results also reveal that the elliptically contoured (EC)

Manuscript received December 17, 2018; revised January 30, 2019 and March 31, 2019; accepted April 17, 2019. Date of publication June 6, 2019; date of current version September 25, 2019. This work was supported in part by the European Research Council (ERC) through the ERC-CoG-Ministry 2014 SEDAL Project under Grant 647423 and in part by the Spanish Ministry of Economy, Industry and Competitiveness under the "Network of Excellence" Program under Grant TEC2016-81900-REDT, Grant TEC2016-77741-R, Grant TIN2015-64210-R, and Grant RTI2018-096765-A-100. The work of J. A. Padrón-Hidalgo was supported by the Grisolia Grant from Generalitat Valenciana (GVA) under Grant GRISOLIA/2016/100. *(Corresponding author: José A. Padrón-Hidalgo.)*

J. A. Padrón-Hidalgo, V. Laparra, and G. Camps-Valls are with the Image Processing Laboratory (IPL), University of Valencia, 46980 Valencia, Spain (e-mail: joseantoniopadronhidalgo@gmail.com; gustau.camps@uv.es).

N. Longbotham is with Descartes Labs, Inc., Santa Fe, NM 87501 USA (e-mail: nathan@descarteslabs.com).

This article has supplementary downloadable material available at http://ieeexplore.ieee.org, provided by the author.

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TGRS.2019.2916212

[1] http://www.esa.int/esaLP/SEMZHM0DU8E_LPgmes_0.html

[2] http://www.nasa.gov/mission_pages/a-train/a-train.html

assumption may be still valid in Hilbert spaces, even when high pervasive distortions mask anomalous targets.

The specific contributions of this paper are as follows.

1) We present an extension of the family of ACD methods presented in [3] to their nonlinear counterparts based on kernel methods. The introduced methods generalize the previous ones and provide more flexible mappings to account for higher order feature dependences.
2) We have tested the robustness of the proposed methods in different scenarios, including simulated, forced, and realistic changes (e.g., floods, droughts, and burned areas). The results of the proposed methods are better than the linear ones in all cases, demonstrating that they can be used in multiple situations. This opens up the option to use the proposed methods not only for the tested situations but also in other problems.
3) We provide a working implementation of all 16 methods as well as a set of labeled images which can be used by other researchers to test ACD methods.

The rest of this paper is outlined as follows. Section II defines the problem and reviews the family of (Gaussian and EC) ACD algorithms introduced in [3]. Section III introduces the proposed kernel-based ACD algorithms. Section IV presents experiments comparing the performance of the proposed algorithms with their linear counterpart in different scenarios. Finally, Section V concludes this paper.

II. Anomalous Change Detection Methods

A. Problem Definition and Literature Review

The change detection (CD) goal is to identify differences in the state of an object, region, or phenomenon by observing it at different temporal times. The CD field is vast and many approaches are available in the literature [2], [8]–[10]. A simple taxonomy could organize them according to three types of products [1], [11]: 1) binary maps; 2) detection of types of changes; and 3) full multiclass change maps, thus including classes of changes and unchanged land-cover classes. Each type of product can be achieved using different sources of information retrieved from the initial spectral images at time instants t_1 and t_2. Unsupervised CD has been widely studied, mainly because it meets the requirements of most applications: 1) the speed in retrieving the change map and 2) the absence of labeled information in applications [2], [12], [13]. However, the lack of labeled information makes the problem of detection more difficult, and thus unsupervised methods typically consider binary change detection problems.

In the last decade, change vector analysis (CVA) techniques have been widely applied: CVA techniques convert the difference image to polar coordinates and operate in such representation space [14], [15]. In [16], morphological operators were successfully applied to increase the discriminative power of the CVA method. In [17], a contextual parcel-based multiscale approach to unsupervised CD was presented. Traditional CVA relies on the experience of the researcher for the threshold definition and is still on-going research [18], [19]. The method has also been studied in terms of sensitivity to differences in registration and other radiometric factors [20]. Another interesting approach based on spectral transforms is the multivariate alteration detection (MAD) [21], [22], where canonical correlation is computed for the points at each time instant and then subtracted. The method consequently reveals changes invariant to linear transformations between the time instants. Radiometric normalization issues for MAD have been recently determined in [23], and nonlinear extensions have also been realized via kernel machines (KMs) [24], [25]. Other approaches based on kernels proposed to use dimensionality reduction via principal components [26] or slow features [27] of the difference image.

A different pathway considers clustering methods. In [28], rather than representing the image difference in the polar domain, local PCAs are used in subblocks of the image, followed by a binary k-means clustering to detect changed/unchanged areas locally. Nonlinear versions of clustering via kernel methods have also been studied. For example, in [29], the kernel k-means parameters were optimized in a fully unsupervised way defining an ANOVA-like cost function. As an alternative to nonlinear kernels, neural networks have also been considered for binary CD [30], [31]. In [32], a Hopfield neural network, where each neuron is connected to a single pixel, is used to enforce neighborhood relationships. Lately, many efforts have been conducted in using deep convolutional neural networks as well [33]–[35].

A related field of investigation in this direction is the so-called ACD [36]. In this field, one looks for changes that are interestingly anomalous in multitemporal series of images and tries to highlight them in contrast to acquisition condition changes, registration noise, or seasonal variation. The interest in ACD is high, and many methods have been proposed in the literature, ranging from regression-based approaches like in the chronocrome [37], where big ("influential") residuals are associated with anomalies [38], [39], to equalization-based approaches that rely on whitening principles [40], as well as multivariate methods [41] that reinforce directions in feature spaces associated with noisy or rare events [21], [42]. The work [43] formalized the field by introducing a framework for ACD, which assumes Gaussianity, yet the derived detector delineates hyperbolic decision functions. Even though the Gaussian assumption reports some advantages (e.g., tractability and generally good performance), it is still an *ad hoc* assumption that it is not necessarily fulfilled in practice. This is the motivation in [3], where the authors introduced EC distributions that generalize the Gaussian distribution and proved more appropriate to modeling fatter tail distributions and thus detect anomalies more effectively. The EC decision functions are pointwise nonlinear and still rely on the second-order feature relations. Recent advances in ACD have considered methods robust to pixel misregistration [44] and sequences of several images [45].

Fig. 1 shows the difference between CD and ACD scenarios using remote sensing images. Changes between two images can be differentiated in regular and anomalous. Regular changes are usually defined by cyclical time patterns, for instance, the change in the vegetation's greenness with the passage of the year's season, exemplified between Fig. 1(a) and (b). On the contrary, an anomalous change is any alteration of the scene that is outside of what is normally

(a) (b) (c) (d)

Fig. 1. Image corresponds to rice farming in the Albufera, Valencia, Spain. (a) Image corresponding to rice crop at planting time. (b) Image corresponding to rice crop at harvest time. (c) Image contains an anomalous change (black square). (d) Image stresses the anomalous location with a green square around it.

expected: for example, the emergence of the black square between Fig. 1(a) and (c). Applying CD and ACD algorithms in Fig. 1(b) and (c) would get similar results. This could not be the case when applied in Fig. 1(a) and (c). On the one hand, the CD algorithm would detect as a change almost all the regions in the image. This would be a good result if one is interested in detecting vegetation changes. However, one could be interested in ignoring the regular changes and detecting only the black square. In such a case, an ACD algorithm would be better fitted since it ignores the brownish to greenish changes and aims to detect as anomaly only the black square.

B. Statistical View of Anomalous Change Detection Problem

Anomalies can be loosely defined as rare items, i.e., with low probability to occur [46], [47]. Also, it is sometimes referred to as outlier, novelty, or extreme detection. An anomalous change is thus a rare, unexpected, change between two consecutive observations (see Fig. 1). In this paper, we want to find samples that can be interpreted as anomalous changes between two multidimensional images. This calls for studying and characterizing differences between multivariate distributions, and in particular, those features that account for the anomalous changes. In [3], a framework to define different anomalous change detectors based on probability distributions was formalized.

Given two images (X and Y), we can treat their pixel values ($\mathbf{x}_i, \mathbf{y}_i$, with $i = 1, \ldots, N$, where N is the number of pixels) as random variables, with probability distributions $\mathbf{x} \sim \mathbb{P}_X$ and $\mathbf{y} \sim \mathbb{P}_Y$, respectively. These distributions can be used to assess how anomalous is in each pixel inside each particular image. On the other hand, let us indicate the joint distribution as $[\mathbf{x}, \mathbf{y}] \sim \mathbb{P}_{X,Y}$, which accounts for how probable particular joint pixel values are, or equivalently to characterize how anomalous a particular change is. For example, if a pixel value changes from $\mathbf{x_i}$ to $\mathbf{y_i}$ and this change has a high probability to occur, it will be classified as a regular change and will not be detected as an anomaly, even if the magnitude change between $\mathbf{x_i}$ and $\mathbf{y_i}$ is highly striking.

The idea is to combine both information to spot only the changes that are not regular. Given two pixels $\mathbf{x}_i, \mathbf{y}_i \in \mathbb{R}^d$ from the same spatial location i but each one from one image, the general formula to compute the amount of anomalousness of a change is

$$\widehat{\mathcal{A}}_{[X,Y]}(\mathbf{x}_i, \mathbf{y}_i) = \frac{\mathbb{P}_X(\mathbf{x}_i)\mathbb{P}_Y(\mathbf{y}_i)}{\mathbb{P}_{[X,Y]}(\mathbf{x}_i, \mathbf{y}_i)}. \tag{1}$$

A sample is detected as anomalous change when it is anomalous with respect to the joint distribution but not anomalous with respect to the distributions of each isolated image. We are using here all three distributions; however, different combinations can be used as we will see in the following.

Instead of using directly (1), it is usual to apply it taking logarithms [3], $\mathcal{A}_{[X,Y]}(\mathbf{x}_i, \mathbf{y}_i) = \log(\widehat{\mathcal{A}}_{[X,Y]}(\mathbf{x}_i, \mathbf{y}_i))$. This can be interpreted in information theoretic terms by noting the relation between probability and information. Elaborating on Shannon's information [48], we have

$$\mathcal{A}_{[X,Y]}(\mathbf{x}_i, \mathbf{y}_i) = I_{[X,Y]}([\mathbf{x}_i, \mathbf{y}_i]) - I_X(\mathbf{x}_i) - I_Y(\mathbf{y}_i)$$

where $I_A(\mathbf{b})$ is the amount of information in Shannon's terms the sample $\mathbf{b}$ provides, assuming that it follows the distribution $\mathbb{P}_A$. In these terms, a sample will be interpreted as an anomalous change if the information obtained by observing the sample in both images simultaneously is big with respect to the information obtained by observing it in each isolated image.

C. Linear ACD Algorithms

Assuming that all three distributions follow a *multivariate Gaussian*, we can express the formula only in terms of covariance matrices. The amount of *anomalousness* is given by

$$\mathcal{A}_{\mathcal{G}}(\mathbf{x}_i, \mathbf{y}_i) = \xi(\mathbf{z}_i) - \beta_x\xi(\mathbf{x}_i) - \beta_y\xi(\mathbf{y}_i) \tag{2}$$

where $\xi(\mathbf{a}) = \mathbf{a}^\top \mathbf{C}_a^{-1}\mathbf{a}$, $\mathbf{C}_a$ is the estimated covariance matrix with the available data, and being $\mathbf{z} = [\mathbf{x}, \mathbf{y}] \in \mathbb{R}^{2d}$. The value of $\beta_x, \beta_y \in \{0, 1\}$ parameters defines which distributions are taken into account to define our anomaly. The different combinations give rise to different anomaly detectors (see Table I). These methods and some variants have been widely used in many hyperspectral image analysis settings because of its simplicity and generally good performance [49]–[51].

However, these methods are hampered by a fundamental problem: the (typically strong) assumption of Gaussianity that is implicit in the formulation. Accommodating other data

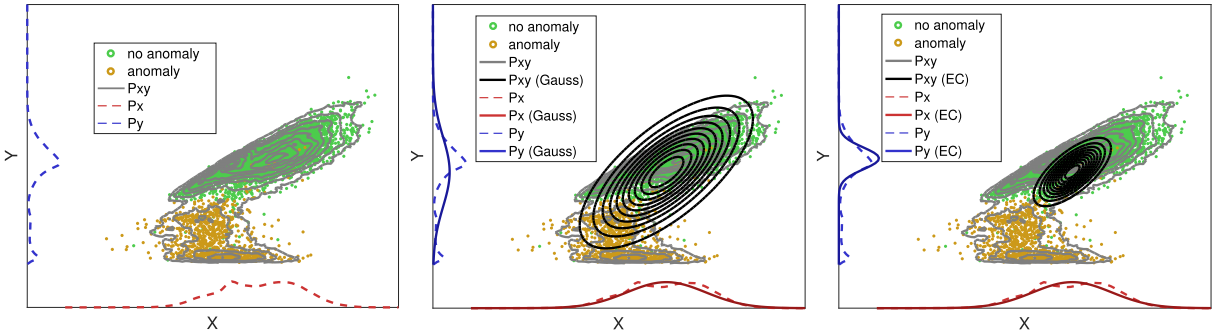


Fig. 2. Illustration of the ACD probabilistic framework. (From left to right) Joint and marginal probability distributions of the original data, the Gaussian model, and the EC model. See text for details.

TABLE I
FAMILY OF ACD ALGORITHMS

ACD algorithm	β_x	β_y
RX	0	0
Chronocrome $\mathbf{y}\|\mathbf{x}$	0	1
Chronocrome $\mathbf{x}\|\mathbf{y}$	1	0
Hyperbolic ACD	1	1

distributions may not be easy in general. Theiler *et al.* [3] introduced alternative ACD to cope with EC distributions [52]. Roughly speaking, the idea is to model the data using an *EC distribution*. EC distributions are particularly convenient in the case of images [53]. In particular, the formulation introduced in [3] uses the multivariate Student's t-distribution, giving rise to the following formula for computing the amount of *EC anomalousness*:

$$\begin{aligned}\mathcal{A}_{\mathrm{EC}}(\mathbf{x}_i,\mathbf{y}_i) = {} & (2d+\nu)\log\left(1+\frac{\xi(\mathbf{z}_i)}{\nu}\right)\\ & -\beta_x(d+\nu)\log\left(1+\frac{\xi(\mathbf{x}_i)}{\nu}\right)\\ & -\beta_y(d+\nu)\log\left(1+\frac{\xi(\mathbf{y}_i)}{\nu}\right) \qquad (3)\end{aligned}$$

where ν controls the shape of the Student's t-distribution: for $\nu \to \infty$, the solution approximates the Gaussian and for $\nu \to 0$, it diverges.

An interesting particular case is the RX algorithm which brings to the same result for the Gaussian and the elliptical case (independently of the ν value). All extra operations applied by the EC formulation with regard to the Gaussian version are increasing monotonic functions, which do not change the ordering of the values. Therefore, although the values of anomalousness are different [i.e. $\mathcal{A}_{\mathcal{G}}(\mathbf{x}_i,\mathbf{y}_i) \neq \mathcal{A}_{\mathrm{EC}}(\mathbf{x}_i,\mathbf{y}_i)$], the values are sorted in the same way which makes the detection curves equal too. The same effect happens between the RX methods based on kernels proposed in Section III.

Fig. 2 shows an example of the distributions involved in the ACD framework. In order to be able to visualize the distributions, we restrict ourselves to the most simple situation, where each image contains just one band. In particular, we show the distributions for band 9 of a Sentinel-2 image over Australia (see Table IV). We show results for the distribution of the data estimated using histograms, when assuming Gaussian or EC distributions. Note that the estimation of the distribution based on histograms is only feasible in the low-dimensional (i.e., 2-D) case: when the number of bands increases, the computation of the histogram becomes unfeasible due to the curse of dimensionality. However, the Gaussian and the EC model can be estimated easily for multiple dimensions. The difference between the Gaussian and the EC model relies on the kurtosis of the distribution; while for the Gaussian model the kurtosis is constant, for the EC model it can be controlled with the ν parameter. By comparing the marginal distributions in the central and the right panels, we can easily spot the differences between the Gaussian and the EC model. For the horizontal axes, the data follow quite well the Gaussian model, i.e., the red solid line and the red dashed line are very similar in the central panel. However, the Gaussian model fails when reproducing the probability for the vertical axes (central panel, blue lines). Although it is not a perfect model, the EC distribution is better suited than the Gaussian distribution for describing the real distribution of the data. For instance, in the case of the P_y (equivalent to $\mathbb{P}_Y$) distribution (blue lines, vertical axes), the EC description (blue solid line in third panel) is more similar to the original one (blue dashed lines) than the description given by the Gaussian distribution (blue solid line in second panel).

III. KERNEL ACD ALGORITHMS

Previous methods are linear and depend on estimating covariance matrices with the available data and use them as a metric for testing anomalousness. These methods are fast to apply and delineate pointwise nonlinear decision boundaries, but still rely on the second-order statistics. This restricts the class of functions that can be implemented and thus the generalization capabilities of the algorithm. For instance, in Fig. 2, the assumed joint distributions (dark green) for both Gaussian and EC models clearly differ from the *r*eal distribution (light green). We here address this issue through the theory of reproducing kernel functions [5], which allows us to capture higher order feature relations while still relying

on linear algebra. Kernel methods are particularly robust to reduced sample sizes and high-dimensional feature spaces, and situations are often encountered in hyperspectral image detection problems.

Kernel methods constitute a well-known approach in machine learning. They have been mainly used for classification and regression, and not that much in anomaly and target detection. The problem has been approached mainly with discriminative and subspace methods: the support vector domain description (SVDD)—also known as one-class SVM—, the kernel OSP, and the kernel RX methods [7]. In our approach, we will proceed with the kernelization of the previous anomaly change detection methods in the same way as for deriving the kernel RX in [54], yet we here extend the framework by assuming EC distributions and parameterizations [see Table I and (3)].

Kernel methods rely on the notion of similarity between points in a higher (possibly infinite) dimensional space. They assume the existence of a Hilbert space $\mathcal{H}$ equipped with an inner product $\langle\cdot,\cdot\rangle_{\mathcal{H}}$. Samples in $\mathcal{X}$ are mapped into $\mathcal{H}$ by means of a feature map $\boldsymbol{\phi}: \mathcal{X} \to \mathcal{H}, \mathbf{x}_i \mapsto \boldsymbol{\phi}(\mathbf{x}_i)$, $1 \leq i \leq n$ [55]. The mapping function $\boldsymbol{\phi}$ can be defined explicitly or implicitly, which is usually the case in the kernel methods. The similarity between the elements in $\mathcal{H}$ can be estimated using its associated inner product $\langle\cdot,\cdot\rangle_{\mathcal{H}}$ via reproducing kernels in Hilbert spaces, $k: \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, such that the pairs of points $(\mathbf{x}_i, \mathbf{x}_j) \mapsto k(\mathbf{x}_i, \mathbf{x}_j)$. So, we can estimate similarities in $\mathcal{H}$ without the explicit definition of the *feature map* $\boldsymbol{\phi}$, and hence without the need of having access to the points in $\mathcal{H}$. The function k is considered a valid *kernel function* if it satisfies Mercer's condition [56].

The mapped training data matrix $\mathbf{X} = [\mathbf{x}_1, \ldots \mathbf{x}_n] \in \mathbb{R}^{n\times d}$ is now denoted as $\boldsymbol{\Phi} = [\boldsymbol{\phi}(\mathbf{x}_1), \ldots \boldsymbol{\phi}(\mathbf{x}_n)] \in \mathbb{R}^{n\times d_{\mathcal{H}}}$. In the following, we show how to estimate the $\xi(\mathbf{x}_i)$ function in the Hilbert space, i.e., $\xi^{\mathcal{H}}(\mathbf{x}_i) = \xi(\boldsymbol{\phi}(\mathbf{x}_i))$. The other terms are derived equivalently. Note that one could think of different mappings for each image, $\boldsymbol{\phi}: \mathbf{x} \to \boldsymbol{\phi}(\mathbf{x})$ and $\boldsymbol{\psi}: \mathbf{y} \to \boldsymbol{\psi}(\mathbf{y})$, $\boldsymbol{\Psi} \in \mathbb{R}^{n\times d_{\mathcal{F}}}$, respectively. However, in our case, we are forced to consider mapping to the same Hilbert space because we have to stack the mapped vectors to estimate $\xi(\boldsymbol{\phi}(\mathbf{z}))$, i.e., $\mathcal{F} = \mathcal{H}$. The mapped training data to Hilbert spaces are denoted as $\boldsymbol{\Phi}$. In order to estimate $\xi(\boldsymbol{\phi}(\mathbf{x}_i))$, we follow the same procedure as in the linear case but first mapping the points to the Hilbert space:

$$\xi^{\mathcal{H}}(\mathbf{x}_i) = \boldsymbol{\phi}(\mathbf{x}_i)(\boldsymbol{\Phi}^\top\boldsymbol{\Phi})^{-1}\boldsymbol{\phi}(\mathbf{x}_i)^\top. \tag{4}$$

Note that we do not have access to either the samples or the covariance in the Hilbert. However, note that $(\boldsymbol{\Phi}^\top\boldsymbol{\Phi})^{-1} = \boldsymbol{\Phi}^\top \quad (\boldsymbol{\Phi}\boldsymbol{\Phi}^\top\boldsymbol{\Phi}\boldsymbol{\Phi}^\top)^{-1}\boldsymbol{\Phi}$. This can be easily shown by right multiplying by the term $\boldsymbol{\Phi}^\top\boldsymbol{\Phi}\boldsymbol{\Phi}^\top$ and applying some linear algebra. By substituting in (4), we get

$$\xi^{\mathcal{H}}(\mathbf{x}_i) = \boldsymbol{\phi}(\mathbf{x}_i)\boldsymbol{\Phi}^\top \quad (\boldsymbol{\Phi}\boldsymbol{\Phi}^\top\boldsymbol{\Phi}\boldsymbol{\Phi}^\top)^{-1}\boldsymbol{\Phi}\boldsymbol{\phi}(\mathbf{x}_i)^\top.$$

In this equation, we can replace all dot products by reproducing kernel functions using the represent theorem [5], and hence

$$\xi^{\mathcal{H}}(\mathbf{x}_i) = \xi(\boldsymbol{\phi}(\mathbf{x}_i)) = \mathbf{k}_i(\mathbf{K}\mathbf{K})^{-1}\mathbf{k}_i^\top$$

where $\mathbf{k}_i \in \mathbb{R}^{1\times n}$ contains the similarities between $\mathbf{x}_i$ and all training data, $\mathbf{X}$, and $\mathbf{K} \in \mathbb{R}^{n\times n}$ stands for the kernel matrix containing all training data similarities. The solution may need extra regularization $\xi^{\mathcal{H}}(\mathbf{x}_i) = \mathbf{k}_i(\mathbf{K}\mathbf{K} + \lambda\mathbf{I}_n)^{-1}\mathbf{k}_i^\top$, $\lambda \in \mathbb{R}^+$. Therefore, the kernel version of ((2)) is

$$\mathcal{A}_G^{\mathcal{H}}(\mathbf{x}_i, \mathbf{y}_i) = \xi^{\mathcal{H}}(\mathbf{z}_i) - \beta_x\xi^{\mathcal{H}}(\mathbf{x}_i) - \beta_y\xi^{\mathcal{H}}(\mathbf{y}_i).$$

By following a similar procedure for (3), one obtains kernel versions of the EC linear solution:

$$\begin{aligned}\mathcal{A}_{\mathrm{EC}}^{\mathcal{H}}(\mathbf{x}_i, \mathbf{y}_i) = {} & (2d+\nu)\log\left(1 + \frac{\xi^{\mathcal{H}}(\mathbf{z}_i)}{\nu}\right) \\ & - \beta_x(d+\nu)\log\left(1 + \frac{\xi^{\mathcal{H}}(\mathbf{x}_i)}{\nu}\right) \\ & - \beta_y(d+\nu)\log\left(1 + \frac{\xi^{\mathcal{H}}(\mathbf{y}_i)}{\nu}\right).\end{aligned}$$

Note that in the case of $\beta_x = \beta_y = 0$, the algorithm reduces to kernel RX which was previously introduced in [54].

Fig. 3 shows the results of different ACD methods for the illustrative example presented in Fig. 2. Different thresholds over the anomalousness function, $\mathcal{A}$, are represented as contour lines. Each method obtains different decision boundaries. The ideal situation would be to have a surface where the green points are surrounded by a contour line and the yellow points are outside of the contour line. Note that this is a complex problem where no perfect solution can be achieved since the anomalous (yellow points) and nonanomalous (green points) pixels are overlapped. Here, and throughout this paper, we will summarize the results using the value of the area under curve (AUC) of the detection receiver operating characteristic (ROC) curves. Bigger AUC means better detection of the anomalous change.

As an illustration, we will take a close look at the results for the method that achieves higher AUC, the K-EC-YX. The shape of the surface tries to keep inside the green points (although some orange points are also included). In general, we can see that the kernel methods obtain better results than their linear counterpart.

Note that the flexibility of the solutions is different for the Gaussian, EC, and the kernel-based methods. The surfaces are direct consequence of the probabilistic model assumed; for instance, in the case of RX for Gaussian and EC assumptions, the surfaces are equivalent to the probabilistic distributions of $\mathbb{P}_{X,Y}$ in Fig. 2. It is clear that the kernel versions have much more capacity to nonlinearly adapt the decision surface to the problem.

IV. Experimental Results

This section analyzes the proposed methods. In order to test the robustness of the results, we perform tests in several simulated and real examples of pervasive and anomalous changes. We evaluate the performance of the methods by using the AUC of ROC curves.

We perform three experiments with data sets with different complexities and controls on the analyzed changes. First,

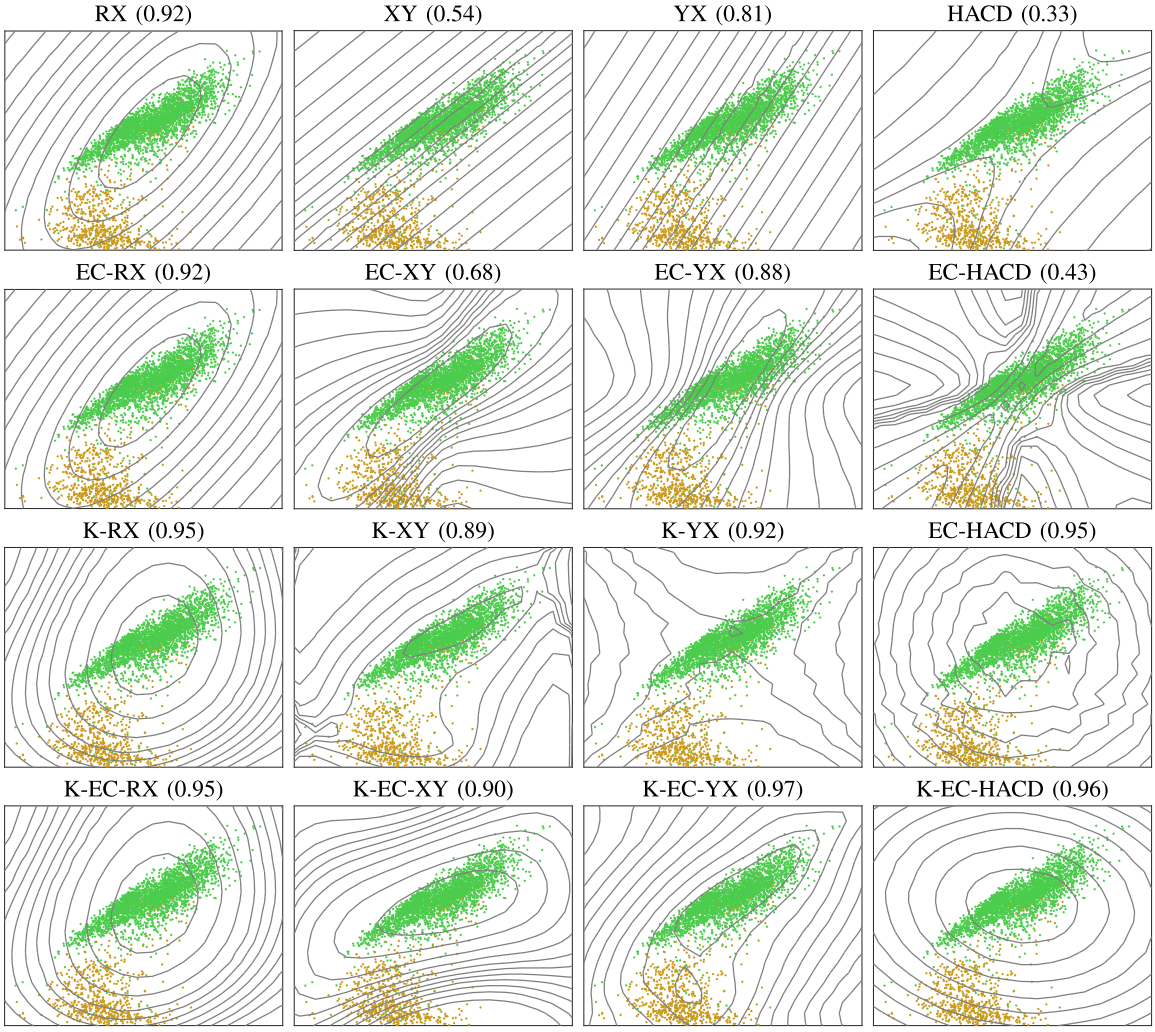


Fig. 3. Toy example of anomalous detection surfaces for different methods, only band 9 of a Sentinel-2 image was employed. Level curves indicate the amount of anomalous (i.e., bigger $\mathcal{A}$). Green dots: nonanomalous data. Yellow points: anomalous data. Overall AUC of the ROC values is given in parenthesis.

we perform an experiment, where we control the anomalous change in a synthetic-controlled scenario. The second experiment deals with data where the changes were real but controlled, since they were manually introduced in the scene using black tarps. Finally, in the third battery of experiments, we deal with natural changes related to floods, droughts, and man-made changes.

We provide MATLAB implementations of the methods. Moreover, we made available a database with the labeled images employed in the third experiment publicly available here: http://isp.uv.es/kacd.html.

A. Experiment 1: Simulated Changes

This experiment is devoted to analyzing the capacity of the methods to detect pervasive and anomalous changes in simulated data by reproducing the simulation framework used in [36]. The data set (see Fig. 4) is an AVIRIS 224-channel image acquired over the Kennedy Space Center (KSC), FL, USA, on March 23, 1996. The data were acquired from an altitude of 20 km and has a spatial resolution of 18 m. After removing low SNR and water absorption bands, a total of 176 bands remain for analysis. More information can be found at http://www.csr.utexas.edu/.

Here, we did not further reduce the dimensionality with PCA and, instead, work directly with the SNR-filtered hyperspectral data. *Pervasive changes* are simulated by adding Gaussian noise with 0 mean and 0.1 standard deviation to all the bands and all the pixels. The image with the added noise is taken as the second image. *Anomalous changes* are produced by scrambling some pixels in the second image. Note that since we are only switching the position of pixels, the global distribution of the image does not change. Since the methods are applied pixelwise, this yields anomalous changes that cannot be detected as anomalies in the individual images.

In this experiment, we restrict ourselves to the use of hyperbolic detectors (HACD), i.e., $\beta_x = \beta_y = 1$, that have shown

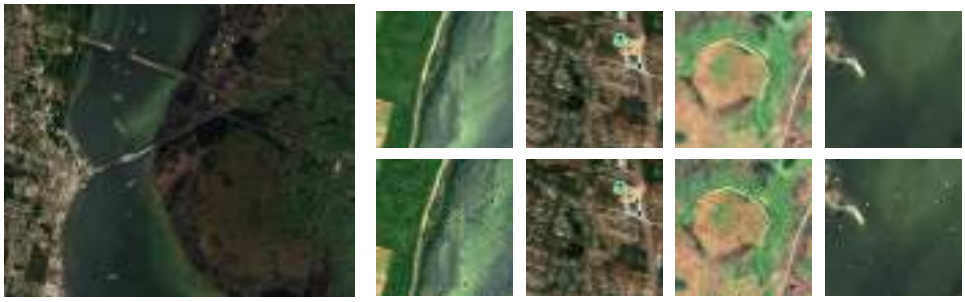


Fig. 4. (Left) AVIRIS hyperspectral image. (Right) Four illustrative chips of simulated changes. (Leftmost) Original image is used to simulate (Rightmost) an anomalous change image by adding Gaussian noise and randomly scrambling 1% of the pixels.

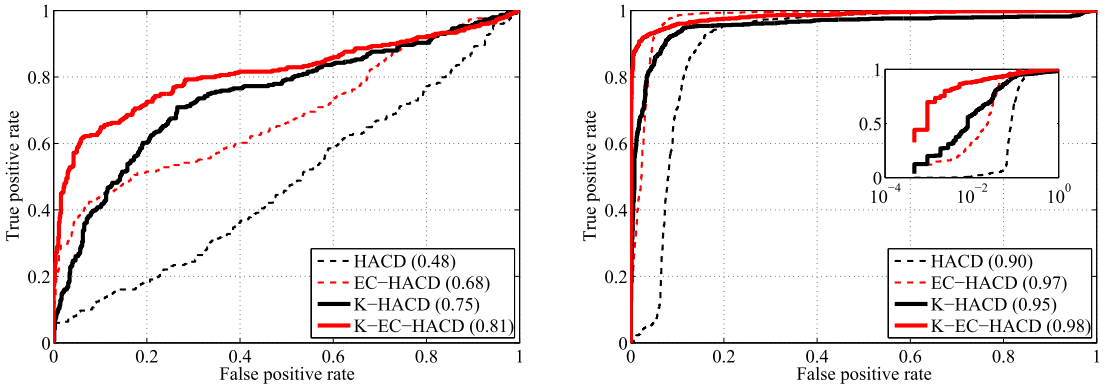


Fig. 5. ROC curves and AUC obtained for simulated changes on hyperspectral image. Results using the HACD detector in this linear (Gaussian and EC) and kernelized version are given. (Left) Results for 100 training examples. (Right) Results for 500 training examples, a version in logarithmic scale is shown in the detailed plot.

improved performance for this particular experiment [3]. We tuned all the involved parameters (estimated covariance $\mathbf{C}_z$ and kernel $\mathbf{K}_z$, ν for the EC methods, and length-scale σ parameter for the kernel versions) through standard cross validation in the training set and show results on the independent test set.

In this experiment, we use the spectral angle mapper (SAM) kernel, $k(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\text{acos}(\mathbf{x}_i^\top \mathbf{x}_j/(\|\mathbf{x}_i\|\|\mathbf{x}_j\|))^2/(2\sigma^2))$, since it has been proven a good choice for hyperspectral images [57]. Two parameters need to be tuned in our kernel versions: the regularization parameter λ and the kernel parameter. In this case, we used $\lambda = 10^{-5}/n$, where n is the number of training samples, and used a isotropic kernel function, whose length-scale σ parameter is tuned in the range of 0.05%–0.95% of the distances between all training samples. We should note that, when a linear kernel is used, $k(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^\top \mathbf{x}_j$, the proposed algorithms reduce to the linear counterparts proposed in [3]. The SAM kernel approximates the linear kernel for high σ values; therefore, results should be improved with regard to the linear versions. Working in the dual (or Q-mode) with the linear kernel instead of the original linear versions can be advantageous *only* in the case of higher dimensionality than available samples, $d \geq n$.

Fig. 5 shows the obtained ROC curves and AUC values for the linear and kernel HACD methods. The data set was split into small training sets of only 100 and 500 pixels, and results are given for 3000 test samples. The main conclusions are that: 1) the kernel versions improve upon their linear counterparts (between 13%–26% in Gaussian and 1%–5% in EC detectors); 2) the EC variants outperform their Gaussian counterparts, especially in the low-sized training sets (+30% over HACD and +18% over EC-HACD in AUC terms); and 3) results improve for all methods when using 500 training samples. The EC-HACD is very competitive compared to the kernel versions in terms of AUC, but still the proposed K-EC-HACD leads to longer tails of false positive detection rates (right figure, inset plot in log-scale).

B. Experiment 2: Real and Enforced Changes

This experiment is designed to analyze the performance of the proposed methods on distortions that are present in real-world imagery. While the distortions that are present in any given pair of image sets are location and sensor-dependent, some of the more prevalent distortions are due to seasonality, look angle, and spatial resolution. These experiments employ a very high spatial resolution sensor that was used to image the same target with highly varying view angles (thus, varying distortion and layover) as well as large differences in seasonality. The ability to detect anomalous changes in these

(a) (b) (c)

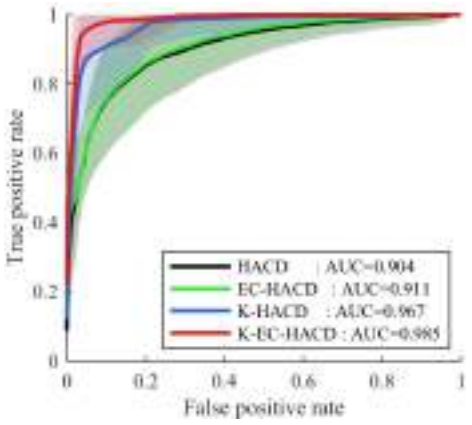
Fig. 6. Three WorldView-2 images present a wide variety of distortions due to both seasonality and view angle. In addition to the more obvious changes in agricultural and natural vegetation, the varying view angles result in variations in ground sample distances (GSD) of 2.0 m (May), 3.6 m (August), and 2.4 m (November). (a) May 2013, 14.0° off nadir. (b) August 2013, 43.6° off nadir. (c) November 2013, 29.3° off nadir.

highly distorted image sets illustrates the unique advantage of these types of algorithms and, in particular, the performance advantages of the proposed methods.

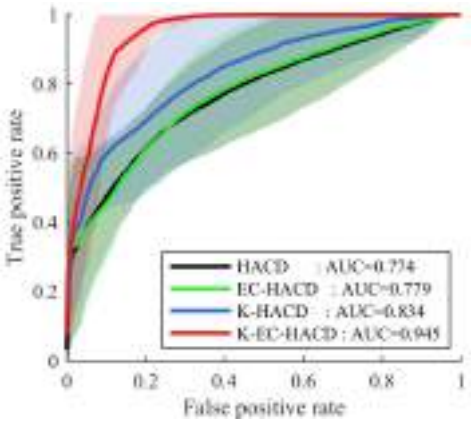
The experiments utilize three WorldView-2 images collected in May, August, and November of 2013. All three images (Fig. 6) were collected over a mixed suburban and rural area with urban residential features, roadways, rivers, and agricultural fields. The first image (May) was acquired at a relatively small off-nadir (14.0°) angle early in the summer season. The second (August) and third (November) images were collected at much higher off-nadir angles, 43.6° and 29.3°, respectively. In each of the final two images, one dark and one white tarp (20 × 20 m each) were introduced as anomalous changes.

This creates two sets on which to test the proposed methods with varying degrees of both angular and seasonality distortions: 1) May/August: high off-nadir difference, moderate seasonality change and 2) May/November: moderate off-nadir difference, large seasonality change. When the white and black tarps that are introduced into the change images are highly anomalous, the spectral change is not unrepresentative of real-world problems. Additionally, the ability to more accurately model changes in highly distorted images provides a unique test case for these proposed methods.

For each experiment, 50 nonanomalous pixels were randomly selected from the stacked image sets to model the data space using the proposed algorithms. 500 randomly selected (training samples held out) nonanomalous and all anomalous pixels (May/August: 153 and May/November: 144) were selected for testing. These random selections were collected for 50 independent runs. The mean ROC curves are reported in Fig. 7 and the statistics for AUC are reported in Table II. As was reported earlier, the parameters ν and σ were tuned through standard cross validation. The results are shown for independent test sets. In both of the experiments, the HACD and EC-HACD methods had almost identical average ROC curves. The parameter search for ν used in the EC-HACD method favored very large values, indicating that the data space is Gaussian and does not particularly benefit from elliptical modeling. This is most likely due to the anomalousness of the tested anomalous targets. Each of the tarp spectral



(a)



(b)

Fig. 7. ROC curves for the two experiments of Section IV-B. The mean value of the experimental runs is plotted with the standard deviation of each detection algorithm represented by the shaded region. (a) May/August: high off-nadir experiment. (b) May/November: large seasonality experiment.

TABLE II

AUC STATISTICS FOR THE WORLDVIEW-2 VIEW-ANGLE AND SEASONALITY EXPERIMENTS

METHODS	May-Aug Large Off-Nadir	May-Nov Large Seasonality
Longmont, Colorado		
HACD	*0.90* ±0.06	*0.77* ±0.08
EC-HACD	*0.91* ±0.06	*0.78* ±0.08
K-HACD	*0.97* ±0.04	*0.83* ±0.11
K-EC-HACD	*0.99* ±0.02	*0.95* ±0.04

signatures is highly anomalous (very dark and very bright), presenting a relatively simplified modeling space. However, the kernel methods did outperform the nonkernel methods by a statistically significant +8% and +17% as measured by mean AUC.

C. Experiment 3: Real and Natural Changes

This experiment deals with the detection of anomalous changes that can be found naturally in a real environment.

1) Data Collection: We collected pairs of multispectral images, and each pair consists of images taken at the same

[3]Only bands in the visible part of the spectrum were used.

TABLE III

IMAGES ATTRIBUTES IN THE EXPERIMENTATION DATA SET

Images	Sensor	Size	Bands	SR
Experiment 1				
KSC	AVIRIS	512 x 614	224	18m
Experiment 2				
Longmont (May)	Worldview-2	1156 x 1563	8	2.0m
Longmont (Aug)	Worldview-2	710 x 1021	8	3.6m
Longmont (Nov)	Worldview-2	1074 x 1149	8	2.4m
Experiment 3				
Argentina	Sentinel-2	1257 x 964	12	10m-60m
Australia	Sentinel-2	1175 x 2031	12	10m-60m
California	Sentinel-2	332 x 964	12	10m-60m
Poopo Lake	MODIS[3]	326 x 201	7	250m-1km
Denver	QuickBird	500 x 684	4	1m-4m

location but at different times. We selected the images in such a way that an anomalous change happened between the two acquisition times. We manually labeled all the images finding the pixels where there is an anomalous change. This step is critical and delicate since we could fall into many false alarms due to, for instance, shadows, illumination changes, or natural changes in the vegetation. This is why this problem is so difficult to solve in an automatic way: for instance, we can see some areas with misclassified pixels in the prediction maps in Fig. 8. All images contain changes of different natures, which allow us to study how the different algorithms perform in a diversity of realistic scenarios. Table III exposes different descriptors of the images in the database. Fig. 8 shows the RGB composites of the pairs of images and the corresponding reference map.

2) Numerical Comparison: Different considerations have to be taken when using the different algorithms. On the one hand, the family of methods based on EC distribution involves the optimization of the ν parameter. On the other hand, kernel methods involve fitting the kernel function parameters. In this experiment, we use the classical RBF kernel which is well suited for multispectral images and has only one parameter, σ. We have performed the experiments using also the SAM and the polynomial kernels; however, results (not shown) were worse than for the RBF kernel. In addition, an extra parameter λ has to be fitted to regularize the matrix inversion. Selecting properly all these three parameters is an issue. An ideal situation would be having a rule of thumb to choose them. We performed preliminary experiments to explore the applicability of several existing rules to estimate the σ parameter. For the different images and problems faced in this section, we applied the heuristics and tried to find an heuristic for the ν and λ parameters. In particular, we investigated ten different heuristics: average distance between all samples, median of the distance between all samples, squared root of the dimensionality times variance per dimension averaged, median of Silverman's rule [58], median of Scott's rule per feature [59], maximum likelihood density estimation, maximum Bayes estimate, maximum entropy estimate, average estimate of marginal kernel density estimate, and kernel density estimation using Gaussian kernel. While some of them have good performance for particular problems, none of the rules was useful in general (results not shown). This is a

TABLE IV

AUC RESULTS FOR ALL FIVE IMAGES. FIRST AND SECOND BEST VALUES FOR EACH IMAGE AND EACH MEMBER OF THE FAMILY ARE IN BOLD. WE PROVIDE THE MEAN AND THE STANDARD DEVIATION FOR TEN DIFFERENT TRIALS, VALUES MARKED WITH (†) HAD AN OUTLIER SO WE GIVE THE MEDIAN INSTEAD OF THE MEAN. VALUES MARKED WITH (•) REPRESENT THE BEST OVERALL RESULT FOR ALL METHODS

METHODS	RX	YX	XY	HACD
ARGENTINA				
ACD	*0.88* ±0.008	*0.86* ±0.010	**0.95** ±0.004	**0.93** ±0.007
K-ACD	**0.93** ±0.009	**0.94** ±0.007	**0.95** ±0.011	**0.93** ±0.005
EC-ACD	*0.88* ±0.008	*0.86* ±0.010	*0.95* ±0.004	*0.93* ±0.006
K-EC-ACD	**0.93** ±0.009	**0.94** ±0.008	• **0.96**±0.008	**0.95** ±0.007
AUSTRALIA				
ACD	*0.79* ±0.019	*0.79* ±0.018	**0.83** ±0.015	*0.79* ±0.012
K-ACD	**0.92** ±0.010	**0.82** ±0.019	**0.83** ±0.049	**0.89** ±0.010
EC-ACD	*0.79* ±0.019	*0.80* ±0.018	*0.83* ±0.015	*0.80* ±0.012
K-EC-ACD	**0.92** ±0.010	**0.86** ±0.016	• **0.95** ±0.008	**0.87** ±0.038
CALIFORNIA (USA)				
ACD	*0.50* ±0.015	*0.59* ±0.017	*0.65* ±0.018	*0.81* ±0.014
K-ACD	**0.61** ±0.024	**0.71** ±0.048	• **0.85** ±0.022	**0.84** ±0.013
EC-ACD	*0.50* ±0.015	*0.59* ±0.016	*0.66* ±0.024	*0.82* ±0.016
K-EC-ACD	**0.61** ±0.024	**0.71** ±0.047	• **0.85** ±0.022	**0.84** ±0.013
DENVER (USA)				
ACD	*0.95* ±0.013	**0.94** ±0.014	*0.82* ±0.059	*0.75* ±0.058
K-ACD	**0.96** ±0.023	†**0.94** ±0.050	**0.87** ±0.017	**0.96** ±0.017
EC-ACD	*0.95* ±0.013	**0.95** ±0.011	*0.88* ±0.027	*0.89* ±0.023
K-EC-ACD	**0.96** ±0.019	†**0.95** ±0.037	• **0.97** ±0.018	• **0.97** ±0.018
POOPO LAKE (BOLIVIA)				
ACD	• **0.99** ±0.002	**0.98** ±0.003	**0.96** ±0.007	*0.63* ±0.032
K-ACD	• **0.99** ±0.002	†*0.97*±0.044	**0.96** ±0.007	**0.96** ±0.005
EC-ACD	• **0.99** ±0.002	**0.98** ±0.004	*0.97* ±0.006	*0.79* ±0.034
K-EC-ACD	• **0.99** ±0.002	**0.98** ±0.013	• **0.99** ±0.002	**0.98** ±0.004

usual problem in ACD where, for instance, instead of setting a particular anomaly threshold, it is usual to compute the ROC curve where all the thresholds are evaluated [3]. Instead of using a different ROC curve for each parameter, we simplified the problem by adopting a cross-validation scheme to fit all the involved parameters: σ, λ, and ν. Note that, not only the kernel methods but also the linear EC methods have hyperparameters to fit. We adopted a realistic scenario where we only need to have labels for a small region. One advantage to use this idea is that once the best parameters are known in a specific region, we can apply this parameter directly without need to use cross validation in similar scenarios. In particular, we use one half of the image for training and obtaining the best parameters, and the other half of the image as the test set. The same procedure was used for all the algorithms.

For each pair of images, we split them into two parts, and we use one for training and one for testing. We select the best parameters by grid search in a cross-validation scheme, using 1000 training samples and 4000 validation samples randomly selected from the training set. Each method implies a different set of parameters. For the ν parameter, we explore 100 points logarithmically spaced between $[10^{-5}, 10^{10}]$. For σ parameter, we explore around the heuristic of the mean of the Euclidean distance between pairs of points (which was the most successful in the preliminary experiments), and we make a grid by taking 60 logarithmically spaced points, respectively, between $[10^{-3}, 10^{3}]$ multiplied by the heuristic value. For the λ parameter, we use 30 values logarithmically spaced between $[10^{-10}, 10^{2.5}]$. Note that these methods do

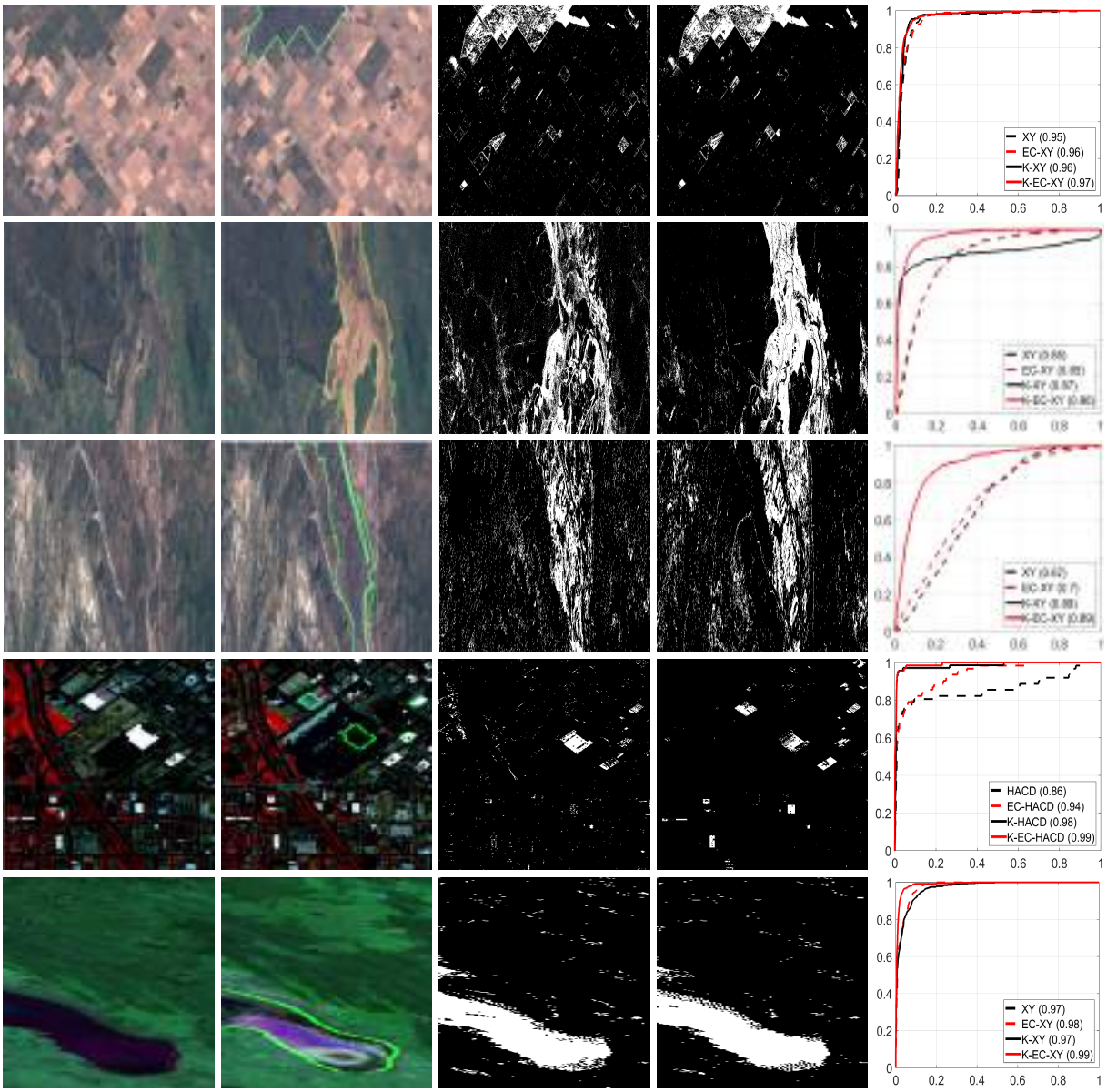


Fig. 8. Images with *natural* anomalous changes, predictions maps, and ROC curves. First row: area burned in Argentina between the months of July and August 2016, and anomalous samples represent 7.5%. Second row: natural floods caused by Cyclone Debbie in Australia 2017, and anomalous samples represent 17.35%. Third row: consequences of the fire in a mountainous area of California, USA, and anomalous samples represent 11.33%. Fourth row: Quickbird multispectral images acquired over Denver, USA, where appears an urbanized area, and anomalous samples represent 1.6%. Last row: drying of Poopo Lake in Bolivia at the end of 2015 and anomalous samples represent 11.7%. First column: images without anomalous changes. Second column: images with anomalous changes and their corresponding labels surrounded with green. Third column: prediction map using the best linear method. Fourth column: prediction map using the best kernel method. Last column: ROC curves and AUC values for the best detectors.

not give a classification but anomalousness value for each pixel. In order to provide a classification map, a particular discrimination threshold (value from which it is decided whether each pixel is an anomalous change or not) should be chosen. It is customary to provide the ROC curves. These curves represent the results of applying a binary classifier to the output of the methods for different threshold values (from more to less restrictive). Each point on the curve is the relationship between true positive and false positive corresponding to the solution provided when applying a particular threshold to the whole data set. ROC analysis is usually employed to compare models. Here, we optimized the parameters of the different methods to maximize the AUC in the training set (top of the image) and used the best parameters for the validation set (bottom of the image).

In Fig. 8, the ROC curves for the best method in AUC terms and Table IV summarizes all AUC values for all images and methods. Fig. 8 compares the ability of the best linear

method against the best kernel method when using the optimal threshold. Moreover, kernel methods produce maps with less false positives and less false negative alarms. As a summary, the kernel version achieves the best results in all the images when compared with its linear counterpart. Although the XY family seems to work better for the K-EC-ACD method of the 16 detectors under study, there is not an overall winner for all the families since each detector has its own characteristics (that can relatively fit data particularities), and the parameters are adjusted according to the type of image. We can see that the K-ACD version obtains a better performance both over the linear ACD and over the linear EC-ACD. And the K-EC-ACD versions have a better performance than the rest. For each type of detector (i.e., RX, XY, YX, or HACD), the AUC values can be ranked as: K-EC-ACD $\geqslant$ K-ACD $\geqslant$ EC-ACD $\geqslant$ ACD.

V. Conclusion

We introduced a family of kernel-based anomaly change detection algorithms. The family extends standard methods such as the RX detector [49], [60] and many others in the literature [3], [43]. The key in the proposed methodology is to redefine the anomaly detection in a reproducing kernel Hilbert space, where the data are mapped to. This endorses the methods with improved capacity and flexibility since nonlinear feature relations (and hence anomalies) can be identified. The introduced methods generalize the previous ones since they account for higher order dependences between features. The proposed methods obtain better results than their linear counterpart for all the performed experiments. We provided implementations of the methods and a database of pairs of images with anomalous changes that can be found in real scenarios http://isp.uv.es/kacd.html.

In practical terms, kernel ACD methods presented here yielded improved results over their linear counterparts in multiple situations. We tested the robustness of this conclusion performing experiments in a wide range of problems. We designed experiments with different complexity levels: synthetic anomaly, real but manually introduced anomaly, and real data where the anomaly has been manually labeled. We analyzed the performance in data coming from different sensors (multispectral and hyperspectral), showing that kernel methods are robust to different numbers of input data dimensions as expected [61]. We adopted standard metrics (AUC and detection) and averaged results over several runs to avoid skewed conclusions.

Interestingly, the EC assumption may still be valid in Hilbert spaces, especially when high pervasive distortions mask anomalous targets. This observation opens the door to the study of the anomalies distribution in Hilbert spaces in the future. A second important conclusion of this paper to be highlighted is that, although the XY family seems to work better for the K-EC-ACD method, among all 16 methods implemented, we did not observe a clear winner in all methods. After all, each problem has its own characteristics and the different methods adapt to different particularities. In the future, we plan to extend the study with low-rank, sparse, and scalable kernel versions to cope with high computational requirements.

References

[1] A. Singh, "Review article digital change detection techniques using remotely-sensed data," *Int. J. Remote Sens.*, vol. 10, no. 6, pp. 989–1003, 1989.

[2] R. J. Radke, S. Andra, O. Al-Kofahi, and B. Roysam, "Image change detection algorithms: A systematic survey," *IEEE Trans. Image Process.*, vol. 14, no. 3, pp. 294–307, Mar. 2005.

[3] J. Theiler, C. Scovel, B. Wohlberg, and B. R. Foy, "Elliptically contoured distributions for anomalous change detection in hyperspectral imagery," *IEEE Geosci. Remote Sens. Lett.*, vol. 7, no. 2, pp. 271–275, Apr. 2010.

[4] N. V. Vapnik, *The Nature of Statistical Learning Theory*, New York, NY, USA: Springer, 1995.

[5] J. Shawe-Taylor and N. Cristianini, *Kernel Methods for Pattern Analysis*, Cambridge, U.K.: Cambridge Univ. Press, 2004.

[6] G. Camps-Valls and L. Bruzzone, *Kernel Methods for Remote Sensing Data Analysis*, Hoboken, NJ, USA:Wiley, 2009.

[7] J. L. Rojo-Álvarez, M. Martínez-Ramón, J. M. Marí, and G. Camps-Valls, *Digital Signal Processing With Kernel Methods*, Hoboken, NJ, USA: Wiley, Feb. 2018.

[8] D. Lu, P. Mausel, E. Brond'izio, and E. Moran, "Change detection techniques," *Int. J. Remote Sens.*, vol. 25, no. 12, pp. 2365–2401, 2004.

[9] D. Manolakis and G. Shaw, "Detection algorithms for hyperspectral imaging applications," *IEEE Signal Process. Mag.*, vol. 19, no. 1, pp. 29–43, Jan. 2002.

[10] D. Manolakis, D. Marden, and A. Gary Shaw, "Hyperspectral image processing for automatic target detection applications," *Lincoln Lab. J.*, vol. 14, no. 1, pp. 79–116, 2003.

[11] P. Coppin, I. Jonckheere, K. Nackaerts, B. Muys, and E. Lambin, "Digital change detection methods in ecosystem monitoring: A review," *Int. J. Remote Sens.*, vol. 25, no. 9, pp. 1565–1596, 2004.

[12] T. Celik, "Unsupervised change detection in satellite images using principal component analysis and k-means clustering," *IEEE Geosci. Remote Sens. Lett.*, vol. 6, no. 4, pp. 772–776, Oct. 2009.

[13] L. Bruzzone and D. F. Prieto, "Automatic analysis of the difference image for unsupervised change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 38, no. 3, pp. 11741–1182, May 2000.

[14] W. A. Malila, "Change vector analysis: An approach for detecting forest change with Landsat," in *Proc. Annu. Symp. Mach. Process. Remotely Sens. Data*, Aug. 1980, pp. 326–336.

[15] F. Bovolo and L. Bruzzone, "A split-based approach to unsupervised change detection in large-size multitemporal images: Application to tsunami-damage assessment," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 6, pp. 1658–1671, Jun. 2007.

[16] M. D. Mura, J. A. Benediktsson, F. Bovolo, and L. Bruzzone, "An unsupervised technique based on morphological filters for change detection in very high resolution images," *IEEE Geosci. Remote Sens. Lett.*, vol. 5, no. 3, pp. 433–437, Jul. 2008.

[17] F. Bovolo, "A multilevel parcel-based approach to change detection in very high resolution multitemporal images," *IEEE Geosci. Remote Sens. Lett.*, vol. 6, no. 1, pp. 33–37, Jan. 2009.

[18] J. Im, J. R. Jensen, and M. E. Hodgson, "Optimizing the binary discriminant function in change detection applications," *Remote Sens. Environ.*, vol. 112, no. 6, pp. 2761–2776, Jun. 2008.

[19] J. Chen, X. Chen, X. Cui, and J. Chen, "Change vector analysis in posterior probability space: A new method for land cover change detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 8, no. 2, pp. 317–321, Mar. 2010.

[20] F. Bovolo, L. Bruzzone, and S. Marchesi, "Analysis and adaptive estimation of the registration noise distribution in multitemporal VHR images," *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 8, pp. 2658–2671, Aug. 2009.

[21] A. A. Nielsen, K. Conradsen, and J. J. Simpson, "Multivariate alteration detection (MAD) and MAF postprocessing in multispectral, bitemporal image data: New approaches to change detection studies," *Remote Sens. Environ.*, vol. 64, no. 1, pp. 1–19, 1998.

[22] A. A. Nielsen, "The regularized iteratively reweighted mad method for change detection in multi-and hyperspectral data," *IEEE Trans. Image Process.*, vol. 16, no. 2, pp. 463–478, Feb. 2007.

[23] M. J. Canty and A. A. Nielsen, "Automatic radiometric normalization of multitemporal satellite imagery with the iteratively re-weighted MAD transformation," *Remote Sens. Environ.*, vol. 112, no. 3, pp. 1025–1036, Mar. 2008.

[24] A. A. Nielsen, "Kernel maximum autocorrelation factor and minimum noise fraction transformations," *IEEE Trans. Image Process.*, vol. 20, no. 3, pp. 612–624, Mar. 2011.
[25] L. Gómez-Chova, R. Santos-Rodríguez, and G. Camps-Valls, "Signal-to-noise ratio in reproducing kernel Hilbert spaces," *Pattern Recognit. Lett.*, vol. 112, no. 1, pp. 75–82, 2018.
[26] M. Ding, Z. Tian, Z. Jin, M. Xu, and C. Cao, "Registration using robust kernel principal component for object-based change detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 7, no. 4, pp. 761–765, Oct. 2010.
[27] C. Wu, L. Zhang, and B. Du, "Kernel slow feature analysis for scene change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 4, pp. 2367–2384, Apr. 2017.
[28] T. Celik, "Multiscale change detection in multitemporal satellite images," *IEEE Geosci. Remote Sens. Lett.*, vol. 6, no. 4, pp. 820–824, Oct. 2009.
[29] M. Volpi, D. Tuia, G. Camps-Valls, and M. Kanevski, "Unsupervised change detection with kernels," *IEEE Geosci. Remote Sens. Lett.*, vol. 9, no. 6, pp. 1026–1030, Nov. 2012.
[30] F. Pacifici, F. del Frate, and W. J. Emery, "Pulse coupled neural networks for detecting urban areas changes at very high resolutions," in *Proc. Joint Urban Remote Sens. Event*, May 2009, pp. 1–7.
[31] F. Pacifici, M. Chini, C. Bignami, S. Stramondo, and W. J. Emery, "Automatic damage detection using pulse-coupled neural networks for the 2009 Italian earthquake," in *Proc. IEEE Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2010, pp. 1996–1999.
[32] S. Ghosh, L. Bruzzone, S. Patra, F. Bovolo, and A. Ghosh, "A context-sensitive technique for unsupervised change detection based on Hopfield-type neural networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 3, pp. 778–789, Mar. 2007.
[33] J. Liu, M. Gong, K. Qin, and P. Zhang, "A deep convolutional coupling network for change detection based on heterogeneous optical and radar images," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 3, pp. 545–559, Mar. 2018.
[34] C. Zhang and Z. Zhang, "Improving multiview face detection with multi-task deep convolutional neural networks," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, Mar. 2014, pp. 1036–1041.
[35] W. Ouyang *et al.*, "DeepID-Net: Object detection with deformable part based convolutional neural networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 7, pp. 1320–1334, Jul. 2017.
[36] J. Theiler, "Quantitative comparison of quadratic covariance-based anomalous change detectors," *Appl. Opt.*, vol. 47, no. 28, pp. F12–F26, 2008.
[37] A. Schaum and A. Stocker, "Long-interval chronochrome target detection," in *Proc. Int. Symp. Spectral Sens. Res.*, 1997, pp. 1760–1770.
[38] D. W. Behnken and N. R. Draper, "Residuals and their variance patterns," *Technometrics*, vol. 14, no. 1, pp. 102–111, 1972.
[39] R. D. Cook, "Detection of influential observation in linear regression," *Technometrics*, vol. 19, no. 1, pp. 15–18, Feb. 1977.
[40] R. Mayer, F. Bucholtz, and D. Scribner, "Object detection by using 'whitening/dewhitening' to transform target signatures in multitemporal hyperspectral and multispectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 41, no. 5, pp. 1136–1142, May 2003.
[41] J. Arenas-García, K. Brandt Petersen, G. Camps-Valls, and L. Kai Hansen, "Kernel multivariate analysis framework for supervised subspace learning: A tutorial on linear and kernel multivariate methods," *IEEE Signal Process. Mag.*, vol. 30, no. 4, pp. 16–29, Jul. 2013.
[42] A. A. Green, M. Berman, P. Switzer, and M. D. Craig, "A transformation for ordering multispectral data in terms of image quality with implications for noise removal," *IEEE Trans. Geosci. Remote Sens.*, vol. 26, no. 1, pp. 65–74, Jan. 1988.
[43] J. Theiler and S. Perkins, "Proposed framework for anomalous change detection," in *Proc. ICML Workshop Mach. Learn. Algorithms Surveill. Event Detection*, 2006, pp. 7–14.
[44] J. Theiler and B. Wohlberg, "Local coregistration adjustment for anomalous change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 50, no. 8, pp. 3107–3116, Aug. 2012.
[45] J. Theiler and S. M. Adler-Golden, "Detection of ephemeral changes in sequences of images," in *Proc. 37th IEEE Appl. Imag. Pattern Recognit. Workshop*, Oct. 2008, pp. 1–8.
[46] Y. Yuan, D. Ma, and Q. Wang, "Hyperspectral anomaly detection by graph pixel selection," *IEEE Trans. Cybern.*, vol. 46, no. 2, pp. 3123–3134, Dec. 2015.
[47] Y. Yuan, Q. Wang, and G. Zhu, "Fast hyperspectral anomaly detection via high-order 2-D crossing filter," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 2, pp. 620–630, Feb. 2016.
[48] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, no. 3, pp. 623–656, Jul./Oct. 1948.
[49] I. S. Reed and X. Yu, "Adaptive multiple-band CFAR detection of an optical pattern with unknown spectral distribution," *IEEE Trans. Acoust., Speech Signal Process.*, vol. 38, no. 10, pp. 1760–1770, Oct. 1990.
[50] C.-I. Chang and S.-S. Chiang, "Anomaly detection and classification for hyperspectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 40, no. 6, pp. 1314–1325, Jun. 2002.
[51] H. Kwon, S. Z. Der, and N. M. Nasrabadi, "Adaptive anomaly detection using subspace separation for hyperspectral images," *Opt. Eng.*, vol. 42, no. 11, pp. 3342–3351, 2003.
[52] S. Cambanis, S. Huang, and G. Simons, "On the theory of elliptically contoured distributions," *J. Multivariate Anal.*, vol. 11, no. 3, pp. 368–385, 1981.
[53] S. Lyu and E. P. Simoncelli, "Nonlinear extraction of independent components of natural images using radial Gaussianization," *Neural Comput.*, vol. 21, no. 6, pp. 1485–1519, Jun. 2009.
[54] H. Kwon and N. M. Nasrabadi, "Kernel RX-algorithm: A nonlinear anomaly detector for hyperspectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 2, pp. 388–397, Feb. 2005.
[55] B. Schölkopf and A. J. Smola, *Learning With Kernels: Support Vector Machines, Regularization, Optimization, and Beyond* (Adaptive Computation and Machine Learning). Cambridge, MA, USA: MIT Press, Dec. 2002, p. 644.
[56] M. A. Aizerman, E. M. Braverman, and L. I. Rozoner, "Theoretical foundations of the potential function method in pattern recognition learning," *Autom. Remote Control*, vol. 25, pp. 821–837, Jun. 1964.
[57] G. Camps-Valls, "Kernel spectral angle mapper," *Electron. Lett.*, vol. 52, no. 14, pp. 1218–1220, 2016.
[58] B. W. Silverman, *Density Estimation for Statistics and Data Analysis*. London, U.K.: Chapman-Hall, 1986.
[59] D. Scott, *Scott's Rule*, vol. 2. Hoboken, NJ, USA: Wiley, 2010.
[60] X. Lu, L. E. Hoff, I. S. Reed, M. Chen, and L. B. Stotts, "Automatic target detection and recognition in multiband imagery: A unified ML detection and estimation approach," *IEEE Trans. Image Process.*, vol. 6, no. 1, pp. 143–156, Jan. 1997.
[61] L. Gömez-Chova, J. Muñoz Marí, V. Laparra, J. Malo-López, and G. Camps-Valls, "A review of kernel methods in remote sensing data analysis," in *Augmented Vision and Reality*, vol. 3, S. Prasad, L. M. Bruce, and J. Chanussot, Eds., New York, NY, USA: Springer, 2011.

José A. Padrón-Hidalgo received the B.Sc. degree in telecommunications and electronics from the University of Pinar del Río, Pinar del Río, Cuba, in 2015. He is currently pursuing the Ph.D. degree in electronics with the Universitat de València, Valencia, Spain.

His research interests include developing algorithms in order to detect anomalous and extreme changes for remote sensing imagery with the Image and Signal Processing Group, Universitat de València.

Valero Laparra was born in Valencia, Spain, in 1983. He received the B.Sc. degree in telecommunications engineering and the B.Sc. degree in electronics engineering from the Universitat de València, Valencia, in 2005 and 2007, respectively, the B.Sc. degree in mathematics from the Universidad Nacional de Educación a Distancia, Madrid, Spain, in 2010, and the Ph.D. degree in computer science and mathematics from the Universitat de València in 2011.

He is currently an Assistant Professor with the Escuela Técnica Superior de Ingeniería, Universitat de València, and a Researcher with the Image Processing Laboratory, Universitat de València.

Nathan Longbotham received the B.S. degree *(magna cum laude)* in physics from Abilene Christian University, Abilene, TX, USA, in 2001, the M.S. degree in optical science and engineering from The University of New Mexico, Albuquerque, NM, USA, in 2008, and the Ph.D. degree in aerospace engineering sciences with a specialization in remote sensing from the University of Colorado, Boulder, CO, USA, in 2012. He is currently a Remote Sensing Scientist with Descartes Labs, Inc., Santa Fe, NM, USA, where he is involved in improving and standardizing data normalization across all sensors served through the Descartes Labs Platform. Descartes is building the missing geospatial link to make satellite imagery useful by providing immediate and convenient access to worldwide imagery through the Descartes Labs Platform. Prior to joining the company, he was involved in a variety of both software- and hardware-based remote sensing technologies, including computational information extraction, tunable ultraviolet lasers for holographic data storage drives, and Q-switched microlasers for LIDAR systems.

Dr. Longbotham is a University Scholar and a Presidential Scholar, and received the Fred J. Barton Departmental Award for his B.S. degree.

Gustau Camps-Valls (M'04–SM'07–F'18) received the Ph.D. degree in physics from the Universitat de València, Valencia, Spain, in 2002.

He is currently a Full Professor of electrical engineering and a Coordinator of the Image and Signal Processing Group, Universitat de València. He is involved in the development of machine learning algorithms for geoscience and remote sensing data analysis. He has authored 200 journal papers, more than 200 conference papers, and 20 international book chapters. He holds a Hirsch's index, h = 60 (source: Google Scholar), entered the ISI list of Highly Cited Researchers in 2011, and Thomson Reuters' ScienceWatch identified one of his papers on Kernel-based analysis of hyperspectral images as a Fast Moving Front research.

Dr. Camps-Valls was a recipient of the Prestigious European Research Council (ERC) Consolidator Grant on Statistical Learning for Earth Observation Data Analysis in 2015. He is/has been an Associate Editor of the IEEE TRANSACTIONS ON SIGNAL PROCESSING, the IEEE GEOSCIENCE AND REMOTE SENSING LETTERS, and the IEEE SIGNAL PROCESSING LETTERS, and an Invited Guest Editor of the IEEE JOURNAL OF SELECTED TOPICS IN SIGNAL PROCESSING in 2012 and the *IEEE Geoscience and Remote Sensing Magazine* in 2015. He serves as an Editor for the books *Kernel Methods Engineering, Signal and Image Processing* (IGI, 2007), *Kernel Methods for Remote Sensing Data Analysis* (Wiley & Sons, 2009), *Remote Sensing Image Processing* (MC, 2011), and *Digital Signal Processing with Kernel Methods* (Wiley & Sons, 2018).

Efficient Nonlinear RX Anomaly Detectors

José A. Padrón Hidalgo, *Student Member, IEEE*, Adrián Pérez-Suay, *Member, IEEE*, Fatih Nar, and Gustau Camps-Valls, *Fellow, IEEE*

***Abstract*— Current anomaly detection (AD) algorithms are typically challenged by either accuracy or efficiency. More accurate nonlinear detectors are typically slow and not scalable. In this letter, we propose two families of techniques to improve the efficiency of the standard kernel Reed–Xiaoli (KRX) method for AD by approximating the kernel function with either the *data-independent* random Fourier features or the *data-dependent* basis with the Nyström approach. We compare all methods for both real multi- and hyperspectral images. We show that the proposed efficient methods have a lower computational cost, and they perform similar to (or outperform) the standard KRX algorithm thanks to their implicit regularization effect. Last but not least, the Nyström approach has an improved power of detection.**

***Index Terms*— Anomaly detection (AD), hyperspectral, kernel methods, low-rank approximation, nonlinear methods, Nyström method, randomization, randomized feature maps, Reed–Xiaoli (RX) detector.**

I. Introduction

ANOMALY detection (AD), as a remote sensing (RS) research topic, is gaining importance because of the need for processing large number of images that are acquired from satellite and airborne platforms [1]. AD aims to detect small portions of the image, which do not belong to the background of the scene. Unlike target detection, AD does not use known target spectra, and anomalies are assumed to be rare and at the tail of the background distribution.

Among the many detector algorithms found in the literature, the Reed–Xiaoli (RX) detector [2] is widely used due to its good performance and simplicity. The RX detector determines the target pixels that are spectrally different from the image background based on the Mahalanobis metric. For the RX to be effective, anomalous targets must be sufficiently small compared with the background and is assumed to follow a Gaussian distribution. However, it has been shown that the Gaussian distribution assumption fails, for example, in the hyperspectral images or with the complex feature relations, especially at the tails of the distribution [3]. As a result, nonlinear versions of the RX have been introduced to mitigate this problem, and the kernel RX (KRX) detector was proposed in [4] to cope with the complex and nonlinear backgrounds. However, the KRX algorithm has not been widely adopted in practice, because, being a kernel method, the memory and computational cost increase, at least quadratically, with the number of pixels. This poses the perennial problem of accuracy versus usability in nonlinear detectors in general and kernel anomaly detectors in particular.

In this letter, we focus on improving the space (memory) and time efficiency of the KRX anomaly detector. Kernel-based anomaly detectors provide excellent detection performance, since they are able to characterize the nonlinear backgrounds [5]. In order to undertake this challenge, we propose to use efficient techniques based on random Fourier features (RFFs) and low-rank approximations (LRXs) to obtain improved performance of the KRX algorithm. We reported our initial efforts using the RFF approach in [6]

In the literature, the *local* and *global* RX-based detectors have been proposed. In local AD [2], pixels in a sliding window are used as input data. Despite their adaptation to local relations, the detection power has been shown to be low recently [3], [7]. Conversely, in global AD, all image pixels are used to estimate the statistics. Thereby, targets with various sizes and shapes can be detected, while the detection of small targets can be difficult. In this letter, all the methods are used in a global setting for the sake of simplicity.

II. RX-Based Anomaly Detection

Among the various AD methods proposed in the literature, one of the most frequently used anomaly detectors is the RX [2]. In this section, we explain the RX method and its kernelized version, the KRX anomaly detector.

A. RX Anomaly Detector

We consider an acquired image reshaped in matrix form as $\mathbf{X} \in \mathbb{R}^{n\times d}$, where n is the number of pixels and d is the total number of channels acquired by the sensor. For simplicity, let us assume that $\mathbf{X}$ is a centered data matrix. The RX detector characterizes the background in terms of the covariance matrix $\mathbf{\Sigma} = 1/d\mathbf{X}^\top\mathbf{X}$. The detector calculates the squared Mahalanobis distance between a test pixel $\mathbf{x}_*$ and the background as follows:

$$D_{\mathrm{RX}}(\mathbf{x}_*) = \mathbf{x}_*^\top \mathbf{\Sigma}^{-1}\mathbf{x}_*. \tag{1}$$

In a global AD setting, as discussed here, $\mathbf{\Sigma}^{-1}$ can be efficiently computed using all the image pixels, since the dimensionality of the image is much lower than the number of

Manuscript received July 3, 2019; revised October 25, 2019 and January 24, 2020; accepted January 25, 2020. This work was supported in part by the European Research Council (ERC) through the ERC-CoG-2014 SEDAL project under Agreement 647423 and in part by the Spanish Ministry of Economy, Industry and Competitiveness through the 'Network of Excellence' Program under Grant TEC2016-81900-REDT. The work of José A. Padrón Hidalgo was supported by the Grisolia Grant from Generalitat Valenciana (GVA) with code GRISOLIA/2016/100. *(Corresponding author: José A. Padrón Hidalgo.)*

José A. Padrón Hidalgo, Adrián Pérez-Suay, and Gustau Camps-Valls are with the Image Processing Laboratory (IPL), Universitat de València, 46010 Valencia, Spain (e-mail: j.antonio.padron@uv.es).

Fatih Nar is with the Department of Computer Engineering, Konya Food and Agriculture University, 42080 Konya, Turkey.

Color versions of one or more of the figures in this letter are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/LGRS.2020.2970582

pixels ($d \ll n$), whereas, in a local AD setting, $\mathbf{\Sigma}_p^{-1}$ needs to be computed for each image pixel p using the centered pixels in a window having an origin at that pixel [3].

B. KRX Anomaly Detector

It is known that a linear RX is computationally efficient and leads to an optimal solution when pixels in $\mathbf{X}$ follow a Gaussian distribution. However, real-life problems are not always Gaussian-distributed, and this requires models that are more flexible. Kernel methods are a possible solution, because they can capture higher order (nonlinear) feature relations, while still using linear algebra operations [5].

In order to develop the KRX, let us consider a mapping for the pixels in the image to a higher dimensional Hilbert feature space $\mathcal{H}$ by means of the feature map $\boldsymbol{\phi} : \mathbf{x} \in \mathbb{R}^d \rightarrow \boldsymbol{\phi}(\mathbf{x}) \in \mathbb{R}^{d_\mathcal{H}}$. The mapped data matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$ is now denoted as $\mathbf{\Phi} \in \mathbb{R}^{n \times d_\mathcal{H}}$. Let us define a kernel function K that, by virtue of the Riesz theorem, can evaluate (reproduce) the dot product between the samples in $\mathcal{H}$, i.e., $K(\mathbf{x}, \mathbf{x}') = \langle \boldsymbol{\phi}(\mathbf{x}), \boldsymbol{\phi}(\mathbf{x}') \rangle \in \mathbb{R}$.

To estimate how anomalous a pixel is using a pixel under test for $\mathbf{x}_* \in \mathbb{R}^d$, we first map $\boldsymbol{\phi}(\mathbf{x}_*)$ and apply the RX formula in (1) as

$$D_{\text{KRX}}(\mathbf{x}_*) = \boldsymbol{\phi}(\mathbf{x}_*)^\top (\mathbf{\Phi}^\top \mathbf{\Phi})^{-1} \boldsymbol{\phi}(\mathbf{x}_*) \tag{2}$$

which, after some linear algebra, can be expressed in terms of kernel matrices [5], [8]

$$D_{\text{KRX}}(\mathbf{x}_*) = \mathbf{k}_*^\top (\mathbf{K}\mathbf{K})^{-1} \mathbf{k}_* \tag{3}$$

where $\mathbf{k}_* = [K(\mathbf{x}_*, \mathbf{x}_1), \ldots, K(\mathbf{x}_*, \mathbf{x}_n)]^\top \in \mathbb{R}^n$ contains the similarities between $\mathbf{x}_*$ and all points in $\mathbf{X}$ using K and $\mathbf{K} \in \mathbb{R}^{n \times n}$ stands for the kernel matrix containing all data similarities [4]. Note that, as in the linear RX method, the KRX also requires centering the data (now in $\mathcal{H}$), which can be easily done.[1] Hereafter, we assume that all kernel matrices are centered.

Note that constructing and inverting a kernel matrix of large n pose a huge computational cost. A simple strategy to alleviate this problem is to draw r samples randomly ($r \ll n$) and use them in the standard KRX, which is here referred to as simple subsampling RX (SRX) and defined as

$$D_{\text{SRX}}(\mathbf{x}_*) = \mathbf{k}_{*:r}^\top (\hat{\mathbf{K}}\hat{\mathbf{K}})^{-1} \mathbf{k}_{*:r} \tag{4}$$

where $\hat{\mathbf{X}} \in \mathbb{R}^{r \times d}$ is a data matrix sampled from $\mathbf{X}$, $\mathbf{k}_{*:r} = [K(\mathbf{x}_*, \mathbf{x}_1), \ldots, K(\mathbf{x}_*, \mathbf{x}_r)]^\top \in \mathbb{R}^r$ contains the similarities between $\mathbf{x}_*$ and $\hat{\mathbf{X}}$, and $\hat{\mathbf{K}} \in \mathbb{R}^{r \times r}$ is a kernel matrix containing data similarities between the points in $\hat{\mathbf{X}}$.

III. Efficient Techniques for KRX

Kernel methods are able to fit nonlinear problems, but they do not scale well when the number of samples grows. We propose using a feature map and LRX approaches to improve the efficiency of the KRX detector. We study the following approximations to the KRX method: RFFs previously studied by Nar *et al.* [6], orthogonal random features (ORF), naive LRX, and Nyström low-rank approximation (NRX).

[1] Centering in feature space can be easily done implicitly by the simple kernel matrix operation $\bar{\mathbf{K}} \leftarrow \mathbf{HKH}$, where $H_{ij} = \delta_{ij} - 1/n$ and δ represents the Kronecker delta $\delta_{i,j} = 1$ if $i = j$ and zero otherwise.

A. Randomized Feature Map Approaches

1) RFFs: An outstanding result in the recent kernel method literature makes use of a classical definition in harmonic analysis to the approximation and scalability [9]. Bochner's theorem states that a continuous shift-invariant kernel $K(\mathbf{x}, \mathbf{x}') = K(\mathbf{x} - \mathbf{x}')$ on $\mathbb{R}^d$ is positive-definite (p.d.) if and only if K is the Fourier transform of a nonnegative measure. If a shift-invariant kernel K is properly scaled, its Fourier transform $p(\mathbf{w})$ is a proper probability distribution. This property is used to approximate the kernel functions with linear projections on a number of D random features as

$$K(\mathbf{x}, \mathbf{x}') \approx \frac{1}{D} \sum_{i=1}^{D} \exp\left(-\mathbb{i}\mathbf{w}_i^\top \mathbf{x}\right) \exp\left(\mathbb{i}\mathbf{w}_i^\top \mathbf{x}'\right)$$

where $\mathbf{w}_i \in \mathbb{R}^d$ are randomly sampled from a data-independent distribution $p(\mathbf{w})$ [9]. Note that we can define a 2-D *randomized* feature map $\mathbf{z} : \mathbb{R}^d \rightarrow \mathbb{R}^{2D}$, which can be *explicitly* constructed as $\mathbf{z}(\mathbf{x}) = (1/\sqrt{2D})[\cos(\mathbf{w}_1^\top \mathbf{x}), \sin(\mathbf{w}_1^\top \mathbf{x}), \ldots, \cos(\sin \mathbf{w}_D^\top \mathbf{x}), \sin(\mathbf{w}_D^\top \mathbf{x})]^\top$ to approximate the radial basis function (RBF) kernel.

Therefore, given n data points (pixels), the kernel matrix $\mathbf{K} \in \mathbb{R}^{n \times n}$ can be approximated with the explicitly mapped data, $\mathbf{Z} = [\mathbf{z}_1 \cdots \mathbf{z}_n]^\top \in \mathbb{R}^{n \times 2D}$, and will be denoted as $\hat{\mathbf{K}} \approx \mathbf{Z}\mathbf{Z}^\top$. However, we do not use such an approach in (3), which would lead to a mere approximation with extra computational cost. Instead, we run the linear RX in (1) with explicitly mapped points onto the RFFs, which reduces to

$$D_{\text{RRX}} = \mathbf{z}_*^\top (\mathbf{Z}^\top \mathbf{Z})^{-1} \mathbf{z}_* \tag{5}$$

and leads to a nonlinear randomized RX (RRX) [6] that approximates the KRX. Essentially, we map the original data $\mathbf{x}_i$ onto a nonlinear space through explicit mapping $\mathbf{z}(\mathbf{x}_i)$ to a $2D$-dimensional space (instead of the potentially infinite feature space with $\boldsymbol{\phi}(\mathbf{x}_i)$) and then use the linear RX formula. This allows us to control the space and time complexity explicitly through D, as one has to store the matrices of $n \times 2D$ and the invert matrices of size $2D \times 2D$ only (see Table I). Typically, parameter D satisfies $D \ll n$ in practical applications.

2) ORFs: An RFF has become a very practical solution for the bottleneck in the kernel methods when n grows. In the RFF, frequencies $\mathbf{w_i}$ are sampled from a particular pdf, and they act as a basis. This, however, may lead to features that are linearly dependent, thus geometrically covering less space. Imposing orthogonality in the basis can be a remedy to this issue, which has led to the ORFs [10]. The linear transformation matrix of the ORF is $\mathbf{W}_{\text{ORF}} = 1/\sigma \mathbf{SQ}$, where $\mathbf{Q}$ is a uniformly distributed random orthogonal matrix. The set of rows of Q forms a basis in $\mathbb{R}^d$. $\mathbf{S}$ is a diagonal matrix, with diagonal entries sampled i.i.d. from the χ-distribution with d degrees of freedom. $\mathbf{S}$ makes the norms of the rows of $\mathbf{SQ}$ and $\mathbf{W}$ (with all the frequencies of RFF) identically distributed. Theoretical results show that ORF achieves a lower error than the RFF for the RBF kernel [10]. This approach follows the above RFF philosophy, and the final anomaly score is now

$$D_{ORX} = \mathbf{z}_*^\top (\mathbf{Z}^\top \mathbf{Z})^{-1} \mathbf{z}_* \tag{6}$$

TABLE I
MEMORY AND TIME COMPLEXITY FOR ALL METHODS

	Space		Time			
Method	T	C^{-1}	T	C	C^{-1}	AD
RX	$-$	d^2	$-$	nd^2	d^3	nd^2
RRX & ORX	nD	D^2	ndD	nD^2	D^3	nD^2
NRX	nr	r^2	ndr	nr^2	r^3	nr^2
KRX	n^2	n^2	n^2d	n^3	n^3	n^3

T is transformation of image into a nonlinear space.
C is matrix (covariance, kernel etc.) and C^{-1} is its inverse.

where each frequency $\mathbf{w}_i$ is a row of $\mathbf{W}_{\text{ORF}}$ and $\mathbf{Z}$ is the matrix formed by the mappings $\mathbf{z}(\mathbf{x}_i)$ of each element in the data set, and $\mathbf{z}_*$ is the mapping of a pixel to be tested.

3) Nyström Approximation: The Nyström method selects a subset of samples to construct an LRX of the kernel matrix [11]. This method approximates the kernel function as $K(\mathbf{x}_*, \mathbf{x}) \approx \mathbf{k}_{*:r}^\top \hat{\mathbf{K}}^{-1} \mathbf{k}_{\mathbf{x}:r}$, where $\mathbf{k}_{\mathbf{x}:r}$ contains the similarities between $\mathbf{x}$ and all r points and $\hat{\mathbf{K}} \in \mathbb{R}^{r\times r}$ stands for the kernel matrix between the points in $\hat{\mathbf{X}}$. Therefore, $\mathbf{k}_*$ can be expressed as

$$\mathbf{k}_* \approx \mathbf{R}^\top \hat{\mathbf{K}}^{-1} \mathbf{k}_{*:r} \tag{7}$$

where $\mathbf{R} \in \mathbb{R}^{r\times n}$ is a matrix that contains similarities between the points in $\hat{\mathbf{X}}$ and the points in $\mathbf{X}$. The similarities were computed using the standard RBF kernel function $k(\mathbf{x}, \mathbf{y}) = \exp(-\|\mathbf{x}-\mathbf{y}\|^2/(2\sigma^2))$.

Using the above definition given in (7), the Nyström method approximates the kernel matrix $\mathbf{K}$

$$\mathbf{K} \approx \mathbf{R}^\top \hat{\mathbf{K}}^{-1} \mathbf{R}. \tag{8}$$

By plugging (7) and (8) into (3), one can define the LRX of the KRX

$$D_{\text{NRX}}(\mathbf{x}_*) = \mathbf{k}_{*:r}^\top \hat{\mathbf{K}}^{-1} \mathbf{R} (\mathbf{R}^\top \mathbf{M} \mathbf{R})^{-1} \mathbf{R}^\top \hat{\mathbf{K}}^{-1} \mathbf{k}_{*:r} \tag{9}$$

where $\mathbf{M} = \hat{\mathbf{K}}^{-1} \mathbf{R}\mathbf{R}^\top \hat{\mathbf{K}}^{-1}$, while $\mathbf{M} \in \mathbb{R}^{r\times r}$. Since $\mathbf{R}$ is not a squared matrix ($r < n$), it is rank-deficient, and we propose to use the pseudoinverse instead of the inverse of $\mathbf{R}^\top \mathbf{M} \mathbf{R}$. By doing this, most of the terms cancel, leading to a more compact equation for the NRX

$$D_{\text{NRX}}(\mathbf{x}_*) = \mathbf{k}_{*:r}^\top (\mathbf{R}\mathbf{R}^\top)^\dagger \mathbf{k}_{*:r}. \tag{10}$$

Note that the NRX involves the inversion of an $r \times r$ matrix, which is much more efficient than the KRX. In addition, the Nyström approach is more generic than using the RFF approaches, as it allows one to approximate all positive-semidefinite kernels, not just shift-invariant kernels. Furthermore, this approximation is data-dependent (i.e., the basis functions are a subset of estimation data itself), which could translate into better results [12].
Reduced-set

4) Connection to Reduced-Set Methods: techniques were successfully used to obtain the sparse kernel methods and LRXs of multivariate kernel methods [13]. This methodology can be applied to approximate the KRX, which leads to (10). In this approach, the data matrix $\mathbf{X} \in \mathbb{R}^{n\times d}$ is subsampled into $\hat{\mathbf{X}} \in \mathbb{R}^{r\times d}$, $r \ll n$, and mapped into $\hat{\mathbf{\Phi}} \in \mathbb{R}^{r\times d_{\mathcal{H}}}$, which, by using (2), leads to obtain the LRX formula

$$D_{\text{LRX}}(\mathbf{x}_*) = \boldsymbol{\phi}(\mathbf{x}_*)^\top \hat{\mathbf{\Phi}}^\top (\hat{\mathbf{\Phi}}\mathbf{\Phi}^\top \mathbf{\Phi}\hat{\mathbf{\Phi}}^\top)^{-1} \hat{\mathbf{\Phi}} \boldsymbol{\phi}(\mathbf{x}_*). \tag{11}$$

Identifying $\mathbf{k}_{*:r} = \hat{\mathbf{\Phi}}\boldsymbol{\phi}(\mathbf{x}_*)$ and $\mathbf{R} = \hat{\mathbf{\Phi}}\mathbf{\Phi}^\top$, (11) leads to

$$D_{\text{LRX}}(\mathbf{x}_*) = \mathbf{k}_{*:r}^\top (\mathbf{R}\mathbf{R}^\top)^{-1} \mathbf{k}_{*:r} \tag{12}$$

which just differs from (10) in the inverse of $\mathbf{R}\mathbf{R}^\top$, and when $\mathbf{R}$ is full rank, they are the same. In the following and in the experiments, we will use only D_{NRX} instead of D_{LRX}, as both are mathematically equivalent.

B. Space and Time Complexity

Table I gives the theoretical computational complexity of the benchmark methods (RX, KRX, and SRX) and proposed methods (RRX, orthogonal RX (ORX), and NRX) presented in this letter. In this letter, we assume $d < D < r \ll n$, since we aim to deal with big data settings. In addition, KRX becomes sufficiently efficient when n is small, e.g., $n < 4000$ for a 200×200 image. As seen in Table I, RX provides the best efficiency; thus, it should be employed for scenes where the data are Gaussian-distributed. However, KRX and the proposed KRX approximations should be used for nonlinear distributions. Clearly, KRX is the least efficient compared with the proposed approximations, and it is also not applicable to big data. Feature map methods, e.g., RRX and ORX, provide the best computational efficiency for the nonlinear (i.e., non-Gaussian) distributions, while the LRX methods, e.g., LRX and NRX, are also efficient yet relatively slower than the feature map methods. Thus, one should choose the proper method based on the image distribution characteristics [14], [15], detection performance requirements, and computational resource limitations. These conclusions are assessed experimentally in the following section.

IV. EXPERIMENTAL RESULTS

This section analyzes the performance of the proposed nonlinear RX AD methods. We performed tests in four real examples and tested the robustness using the area under curve (AUC) of the receiver operating characteristic (ROC) curves. We provide an illustrative source code for all methods in http://isp.uv.es/code/fastrx.html.

A. Data Collection and Experimental Setup

We collected multispectral and hyperspectral images acquired by the Quickbird and AVIRIS sensors. Fig. 1 showcases the scenes used in the experiments. The AD scenarios consider anomalies related to latent fires, vehicles, urbanization (roofs), and ships [7], [16], [17]. Table II summaries the relevant attributes of the data sets such as sensors, and spatial and spectral resolution.

Parameter estimation is required for the RX, KRX, RRX, ORX, and NRX. First, the KRX method and its proposed variants involve the optimization of the σ parameter of the RBF kernel. For the feature map approaches (RRX and ORX), the number of basis, D, parameter should be optimized.

(a) (b) (c) (d)

Fig. 1. Images with anomalies (outlined in yellow) in four scenarios. (a) Consequences of the hot spots corresponding to latent fires at the World Trade Center (WTC) in NYC (extension of anomalous pixels represents the 0.23% of the image). (b) Urban area where anomalies are vehicles in Gainesville city (0.52%). (c) Quickbird multispectral images acquired over Denver; the anomalies are roofs in an urbanized area (1.6%). (d) Beach scene where the anomalies are ships captured by the AVIRIS sensor (2.02%) over San Diego, USA.

TABLE II
IMAGE ATTRIBUTES USED IN THE EXPERIMENTATION DATA SET

Images	Sensor	Size	Bands	Resolution
WTC	AVIRIS	200 x 200	224	1.7 m
Gainesville	AVIRIS	100 x 100	190	3.5 m
Denver	Quickbird	500 x 684	4	1m-4m
San Diego	AVIRIS	100 x 100	193	7.5 m

Whereas, for NRX, the number of random subsamples, r, parameter should be optimized.

We adopted a cross-validation scheme to select all the involved parameters: number of Fourier basis D, rank r, and RBF parameter σ. We selected the parameters using different data sizes ranging between 10^3 and 3×10^4 samples.

B. Numerical Comparison

We report the averaged AUC results for all cases with 1000 runs (standard deviations were always lower than 3×10^{-3} and, hence, are not reported). Fig. 2 shows that nonlinear methods improve detection over the linear RX, and the NRX outperforms the other approximations in three out of the four images. The AUC values of the KRX are related to the inversion of a relatively big matrix. This raises the issues of poorly estimated matrices (with a huge condition number), which are also computationally expensive to invert [$\mathcal{O}(n^3)$]. However, all the proposed fast KRX methods have the advantage of solving both issues. First, thanks to the cross-validation procedure, an estimate of the optimal number of features (RRX, ORX) or samples (NRX) can be obtained, allowing to better capture the intrinsic dimensionality of the mapped data. In a previous work [18], we showed that optimizing the number of frequencies in the RFF approaches acts as an efficient regularizer, leading to better estimates with a reduced number of frequencies needed. Second, fast versions are able to obtain better performance in the AUC metric at a fraction of cost (see Fig. 2).

C. On the Computational Efficiency

Fig. 3 illustrates the tradeoff between the computational execution time and the AUC. The crosses indicate different values of rank (D or r parameters) in the set $\{50, 100, 200, 400, 500\}$, and the number of pixels was fixed to $n = 3000$. The optimal parameters estimated for KRX are used for the fast approaches.

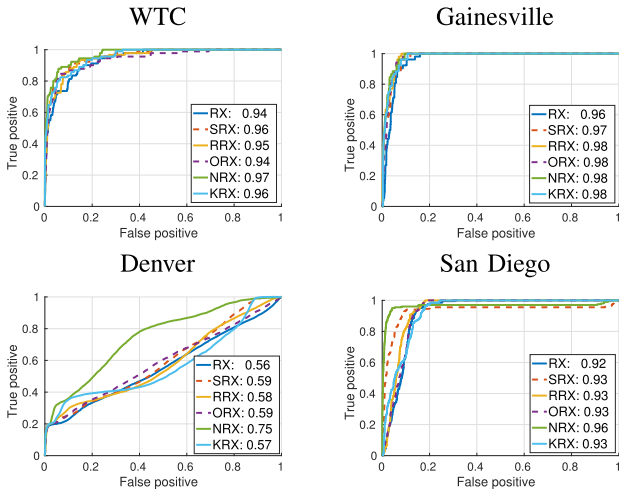


Fig. 2. ROC curves in linear scale for all scenes. Numbers in legend display the AUC values for each method.

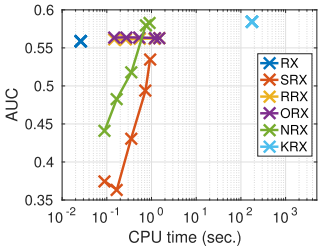


Fig. 3. CPU execution time versus the AUC values for $n = 3000$ pixels; crosses corresponds to different rank values for the Denver image.

The KRX has the best AUC values in all the images. NRX and SRX are more sensitive to the rank values. The RRX and ORX are almost insensitive to the rank, but results do not improve when the rank increases, thus limiting their performance. The combination of lower spectral information and the ambiguity of the class (note that the anomaly class "urbanized" can be confused with a pervasive class "urban") makes the Quickbird scene a very difficult problem (lower AUCs). In this situation, as the rank parameter r for the NRX method grows, it approximates the KRX algorithm. In Fig. 4, the RX detector (top row) is shown against the best detector obtained (bottom row). The best result in the AUC was achieved by the NRX in all the images. It is worth mentioning the good results in detection achieved by the NRX

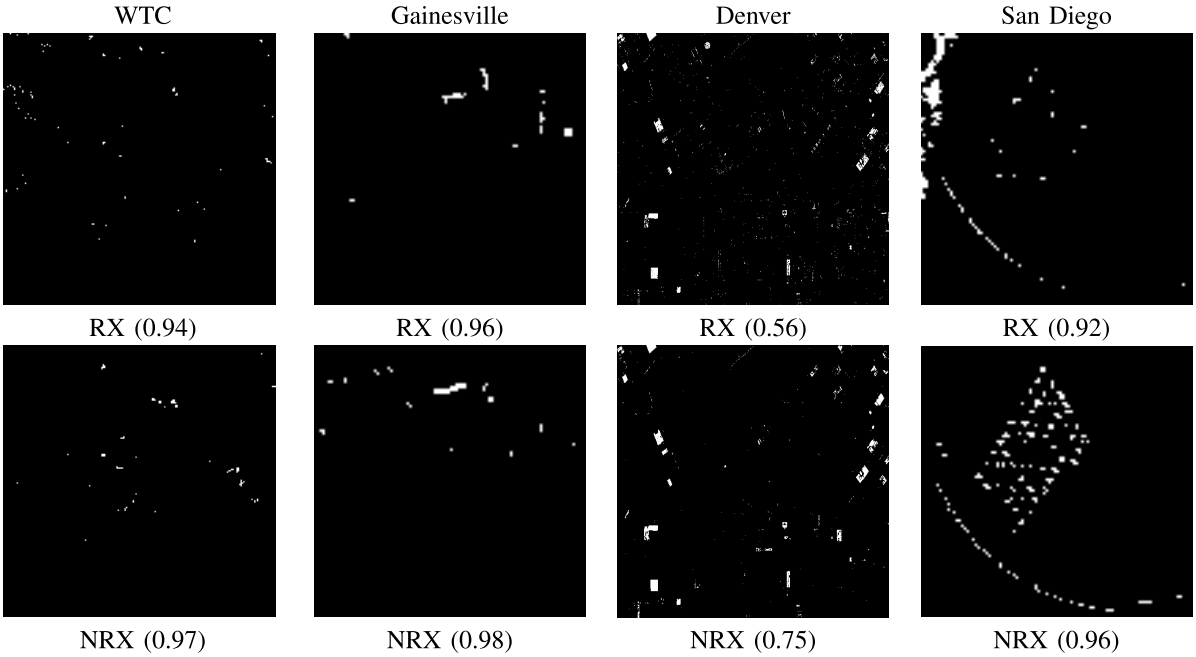


Fig. 4. AD maps for best thresholds. (Top) Best linear RX (AUC) results. (Bottom) Best nonlinear RX (AUC) method).

in all the scenes, which can be visually compared with the linear version.

V. Conclusion

In this letter, we introduced a family of efficient nonlinear AD algorithms based on the RX method. We used the theory of reproducing kernels and proposed several efficient methods. The KRX detector was improved using efficient and fast techniques based on feature maps and LRXs. Among all methods, both the Nyström and the equivalent low-rank (LRX) approximation achieve the best results and yield a more efficient and accurate nonlinear RX method to be applied in practice. For future research, we plan to study the behavior of fast approximations for alternative KRX variants [19], [20]. Note that the presented methodologies for fast KRX can be applicable to other kernel anomaly detectors, in local settings, and for real-time detection.

References

[1] D. W. J. Stein, S. G. Beaven, L. E. Hoff, E. M. Winter, A. P. Schaum, and A. D. Stocker, "Anomaly detection from hyperspectral imagery," *IEEE Signal Process. Mag.*, vol. 19, no. 1, pp. 58–69, Jan. 2002.

[2] I. Reed and X. Yu, "Adaptive multiple-band CFAR detection of an optical pattern with unknown spectral distribution," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 38, no. 10, pp. 1760–1770, 1990.

[3] S. Matteoli, M. Diani, and G. Corsini, "A tutorial overview of anomaly detection in hyperspectral images," *IEEE Aerosp. Electron. Syst. Mag.*, vol. 25, no. 7, pp. 5–28, Jul. 2010.

[4] H. Kwon and N. Nasrabadi, "Kernel RX-algorithm: A nonlinear anomaly detector for hyperspectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 2, pp. 388–397, Feb. 2005.

[5] G. Camps-Valls and L. Bruzzone, *Kernel Methods for Remote Sensing Data Analysis*. London, U.K.: Wiley, 2009.

[6] F. Nar, A. Pérez-Suay, J. A. Padrón-Hidalgo, and G. Camps-Valls, "Randomized RX for target detection," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2018, pp. 4237–4240.

[7] Q. Guo, R. Pu, and J. Cheng, "Anomaly detection from hyperspectral remote sensing imagery," *Geosciences*, vol. 6, no. 4, p. 56, Dec. 2016.

[8] H. Kwon and N. Nasrabadi, "A comparative analysis of kernel subspace target detectors for hyperspectral imagery," *EURASIP J. Adv. Signal Process.*, vol. 2007, Dec. 2007, Art. no. 29250.

[9] A. Rahimi and B. Recht, "Random features for large-scale kernel machines," in *Proc. Neural Inf. Process. Syst. (NIPS)*, 2008, pp. 1177–1184.

[10] F. X. X. Yu, A. T. Suresh, K. M. Choromanski, D. N. Holtmann-Rice, and S. Kumar, "Orthogonal random features," in *Proc. Neural Inf. Process. Syst. (NIPS)*, 2016, pp. 1975–1983.

[11] C. K. I. Williams and M. Seeger, "Using the Nyström method to speed up kernel machines," in *Proc. Neural Inf. Process. System (NIPS)*, 2001, pp. 682–688.

[12] T. Yang, Y.-F. Li, M. Mahdavi, R. Jin, and Z.-H. Zhou, "Nyström method vs random Fourier features: A theoretical and empirical comparison," in *Proc. Neural Inf. Process. Syst. (NIPS)*, Curran Associates, Inc., 2012, pp. 476–484.

[13] J. Arenas-Garcia, K. B. Petersen, G. Camps-Valls, and L. K. Hansen, "Kernel multivariate analysis framework for supervised subspace learning: A tutorial on linear and kernel multivariate methods," *IEEE Signal Process. Mag.*, vol. 30, no. 4, pp. 16–29, Jul. 2013.

[14] D. Manolakis, L. G. Jairam, D. Zhang, and M. Rossacci, "Statistical models for LWIR hyperspectral backgrounds and their applications in chemical agent detection," *Proc. SPIE*, vol. 6565, May 2007, Art. no. 656525.

[15] N. Keshava, "Distance metrics and band selection in hyperspectral processing with applications to material identification and spectral libraries," *IEEE Trans. Geosci. Remote Sens.*, vol. 42, no. 7, pp. 1552–1565, Jul. 2004.

[16] X. Kang, X. Zhang, S. Li, K. Li, J. Li, and J. A. Benediktsson, "Hyperspectral anomaly detection with attribute and edge-preserving filters," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 10, pp. 5600–5611, Oct. 2017.

[17] J. A. Padron-Hidalgo, V. Laparra, N. Longbotham, and G. Camps-Valls, "Kernel anomalous change detection for remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 10, pp. 7743–7755, Oct. 2019.

[18] P. Morales-Alvarez, A. Perez-Suay, R. Molina, and G. Camps-Valls, "Remote sensing image classification with large-scale Gaussian processes," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 2, pp. 1103–1114, Feb. 2018.

[19] J. Theiler and G. Grosklos, "Problematic projection to the in-sample subspace for a kernelized anomaly detector," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 4, pp. 485–489, Apr. 2016.

[20] J. Theiler and G. Grosklos, "Cracks in KRX: When more distant points are less anomalous," in *Proc. 8th Workshop Hyperspectral Image Signal Process., Evol. Remote Sens. (WHISPERS)*, Aug. 2016, pp. 1–5.

Unsupervised Anomaly and Change Detection with Multivariate Gaussianization

José A. Padrón-Hidalgo, Valero Laparra, and Gustau Camps-Valls, *Fellow, IEEE*

***Abstract*—Anomaly detection is a field of intense research in remote sensing image processing. Identifying low probability events in remote sensing images is a challenging problem given the high-dimensionality of the data, especially when no (or little) information about the anomaly is available a priori. While plenty of methods are available, the vast majority of them do not scale well to large datasets and require the choice of some (very often critical) hyperparameters. Therefore, unsupervised and computationally efficient detection methods become strictly necessary, especially now with the data deluge problem. In this paper, we propose an unsupervised method for detecting anomalies and changes in remote sensing images by means of a multivariate Gaussianization methodology that allows to estimate multivariate densities accurately, a long-standing problem in statistics and machine learning. The methodology transforms arbitrarily complex multivariate data into a multivariate Gaussian distribution. Since the transformation is differentiable, by applying the change of variables formula one can estimate the probability at any point of the original domain. The assumption is straightforward: pixels with low estimated probability are considered anomalies. Our method is flexible enough to describe any multivariate distribution, makes an efficient use of memory and computational resources, and is parameter-free. We show the efficiency of the method in experiments involving both anomaly detection and change detection in different remote sensing image sets. For anomaly detection we propose two approaches. The first using directly the Gaussianization transform and the second using an hybrid model that combines Gaussianization and the Reed-Xiaoli (RX) method typically used in anomaly detection. Results show that our approach outperforms other linear and nonlinear methods in terms of detection power in both anomaly and change detection scenarios, showing robustness and scalability to dimensionality and sample sizes.**

***Index Terms*—Change Detection (CD), Anomaly detection, Extremes, Gaussianization, principal component analysis, information, deep learning, probability density estimation.**

I. Introduction

Remote Sensing (RS) has become a powerful tool to develop applications for Earth monitoring [1]–[3]. Earth observation (EO) satellite missions, such as Sentinels-2 and Landsat-8 are able to replace the hard and costly work of the man on the ground. Also, the use of very high resolution (VHR) satellite imagery (e.g. QuickBird and the Worldview constellation) is becoming increasingly important for remote sensing applications, and it makes possible the detection of dangerous events such as extreme precipitations, heat waves, latent fires, droughts, floods or urbanization. The vast amount of data available from different sensors makes it urgent to have automatic methods to detect these events. A good and quite standard approach nowadays to tackle this problem considers statistical models that allow us to detect anomalies and changes on the Earth cover.

Statistical methods for Anomaly detection (AD) focus on detecting small portions of the image which do not belong to the background of the scene [4]. Anomalies are considered a group of weird (low probability) pixels which significantly differ from their neighbors. AD is a challenging task and many variants have been proposed in the literature, such as neighbor based, clustering, classification, etc [5]–[7]. However, among all of them, the Reed-Xiaoli (RX) approach [8] is still the most widely used method for AD since the Gaussian distribution assumption is a reasonable approach in several cases, it is unsupervised, fast and easy to implement. The RX method allows us to detect the anomalous samples compared to background using the well-known Mahalanobis distance. Nevertheless, since the Gaussian assumption is not flexible enough in most cases, variants of the RX has been developed to cope with higher-order feature relations. One option which obtains good results is based on the theory of reproducing kernels in Hilbert spaces, which extend the RX approach to the kernel RX (KRX) [9], [10]. However, the KRX algorithm has not been widely adopted in practice because, being a kernel method, the memory and computational cost increase with the number of pixels cubically and quadratically respectively, and more importantly the selection of the kernel parameters is critical to achieve a good performance. While unsupervised approaches to fit the kernel parameter exist, they achieve a sub-optimal performance and hence supervised approaches have to be used. In this manuscript we propose to use a different, more straigthforward approach to the problem of anomaly detection based on multivariate Gaussianization transformation. The proposed method is able to handle multidimensional data and at the same time does not require additional information to fit any parameter [11]. In order to control the flexibility of the method we propose the combination of the Gaussian assumption (RX) and the Gaussianization transformation, which leads to a powerful, automatic, unsupervised, algorithm for anomaly detection.

Change detection (CD) can be consider a particular case of the anomaly detection problem, where the change class is the target class to be detected. Detecting changes in images

Manuscript received November 14, 2020.

Image Processing Laboratory (IPL)
Universitat de València, Catedrático A. Escardino - 46980 Paterna, València (iSpain). E-mail: gustau.camps@uv.es

Research funded by the European Research Council (ERC) under the ERC-CoG-2014 SEDAL project (grant agreement 647423) and the Spanish Ministry of Economy, Industry and Competitiveness under the 'Network of Excellence' program (grant code TEC2016-81900-REDT). Jose A. Padrón was supported by the Grisolia grant from Generalitat Valenciana (GVA) with code GRISOLIA/2016/100.

automatically is extremely important because it allows us to improve predictions and our understanding of events occurring over the entire surface of the Earth. As for AD, a similar problem occurs when using statistical methods for CD [12]: one aims to learn the distribution of the original image and analyze the statistical differences of the pixels in the new incoming image. Likewise, the RX and KRX extensions have been proposed to deal with CD problems. However, they show the same drawbacks as in AD. Here we describe how the multivariate Gaussianization can be used also in CD problems. Note that in both cases, AD and CD, the proposed statistical method is used to evaluate the pixel's probability so that one can classify them as anomalies or changes, respectively.

The remainder of the manuscript is organized as follow. Section II summarizes the Gaussianization transformation in general, and how to adapt it to anomaly and change detection. In section III, we illustrate its performance in three experiments, involving simulated anomalies, an real AD and CD examples with a database of real multi- and hyperspectral images. Results show that the proposed approach is robust and flexible enough to be applied in different AD and CD scenarios, and obtains better performance than other simple and robust methods (like the RX) and more flexible and adaptable ones, like the KRX. Section IV concludes the paper with some remarks and further work.

II. Multivariate Gaussianization for Detection

The rotation-based iterative Gaussianization (RBIG) is a nonparametric method for density estimation of multivariate distributions [11]. RBIG is rooted in the idea of Gaussianization, introduced in the seminal work by [13] and further developed in [11], [14], which consists of seeking for a transformation G_x that converts a multivariate dataset $\mathbf{X} \in \mathbb{R}^{\ell \times d}$ in domain X to a domain where the mapped data $\mathbf{Y} \in \mathbb{R}^{\ell \times d}$ follows a multivariate normal distribution in domain Y, i.e. $p_Y(\mathbf{y}) \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$:

$$\begin{array}{rccc} G_x: & \mathbf{x} \in \mathbb{R}^d & \mapsto & \mathbf{y} \in \mathbb{R}^d \\ & \sim p_X(\mathbf{x}) & & p_Y(\mathbf{y}) \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d), \end{array} \tag{1}$$

where inputs and mapped data points have the same dimensionality $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, $\mathbf{0}$ is a vector of zeros (for the means) and $\mathbf{I}_d$ is the identity matrix for the covariance of dimension d. Using the change of variable formula one can estimate the probability of a point $\mathbf{x}$ in the original domain:

$$p_X(\mathbf{x}) = p_Y(\mathbf{y}) |J_{G_x}(\mathbf{y})|, \tag{2}$$

where $p_X(\mathbf{x})$ is the probability distribution of the original data point $\mathbf{x}$, and $|J_f(\mathbf{y})|$ is the determinant of the Jacobian of the transformation G_x in the point $\mathbf{y}$. For this formula to work, G_x has to be differentiable, i.e. the $|J_{G_x}(\mathbf{y})| > 0, \forall \mathbf{y}$. The Gaussianization method we propose in this paper, RBIG, obtains a transformation G_x that fulfills this property, cf. [11]. The other part of the product is easy to compute since $p_Y(\mathbf{y})$ can be estimated since p_Y is a multivariate Gaussian by construction. Therefore RBIG can be easily applied to estimate the probability of data points in the original domain, $p_X(\mathbf{x})$.

RBIG is an iterative algorithm, where in each iteration, n, two steps are applied: 1) a set of d marginal Gaussianizations to each of the variables, $\mathbf{\Psi} = [\Phi_1, \ldots, \Phi_d]$, and 2) a linear rotation, $\mathbf{R} \in \mathbb{R}^{d \times d}$:

$$\mathbf{x}[n+1] = \mathbf{R}[n] \cdot \mathbf{\Psi}[n](\mathbf{x}[n]), \quad n = 1, \ldots, N \tag{3}$$

where N is the number of steps (iterations) in the sequence, $n = 1, \ldots, N$. The final transformation G_x is the composition of all performed transformations through iterations. In [11] we showed that with enough iterations the method converges and the transformed data follows finally a standardized Gaussian, i.e. $p_Y(\mathbf{y}) \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$, taking $\mathbf{y} = \mathbf{x}[N]$.

An illustration of how RBIG can be adapted to describe the distribution of remote sensing data is shown in Fig. 1. In this example we take data from the Sentinel-2b image Australia (see Table I for details), which has $d = 12$ bands, and use RBIG to Gaussianize its pixel's distribution. We can see that the Gaussianized data follows a Gaussian distribution. Besides we apply the inverse of the learned Gaussianization transformation to randomly generated Gaussian points obtaining synthetic new data that follows a deemed similar distribution as the original one. This illustrates the invertibility property of RBIG, which allows us to estimate densities in the original domain and use the well-known relation between probability and anomaly to derive unsupervised density-based anomaly and change detectors.

A. RBIG for Detection of Anomalies

One of the most successful methods applied to the problem of anomaly detection is the Reed-Xiaoli (RX) method [8], a successful type of matched filter. The idea behind the RX method can be interpreted in probabilistic terms [10]; intuitively, a data point is more anomalous when it has less probability to appear:

$$A(\mathbf{x}) \propto \frac{1}{p_X(\mathbf{x})}. \tag{4}$$

Actually, when the distribution is assumed to be Gaussian, $p_X \sim p_G$, this relation defines the RX method anomaly detector, i.e. $A(\mathbf{x}) \sim A_{\text{RX}}(\mathbf{x})$. Actually $A_{\text{RX}}(\mathbf{x})$ is equivalent to the Mahalanobis distance between the data point and the mean, i.e. $A_{\text{RX}}(\mathbf{x}) = (\mathbf{x} - \mu)^\top \mathbf{\Sigma}^{-1} (\mathbf{x} - \mu)$, where $p(\mathbf{x}) \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$.

While RX has been widely used, it has the limitations inherent to the Gaussian distribution assumption. The use of kernel methods has been proposed to generalize the RX method to the nonlinear and non-Gaussian case [9], [10]. Kernel methods define the covariance in a higher dimensional Hilbert feature space, which in the RX method translates into replacing the covariance matrix by a kernel matrix that estimates the similarity between samples [15], [16]. In practice this implies that correlation is substituted by a non-linear similarity measure. Therefore the anomaly detected using the kernel RX (KRX) method can be formulated as:

$$A_{\text{KRX}}(\mathbf{x}) \propto \frac{1}{p_K(\mathbf{x})}, \tag{5}$$

where $p_K(\mathbf{x})$ is the distribution induced by using the kernel function instead of the covariance. The kernel RX (KRX) is an elegant extension of the RX, yet it has the problem of fitting kernel parameters and the high computational cost (as one has

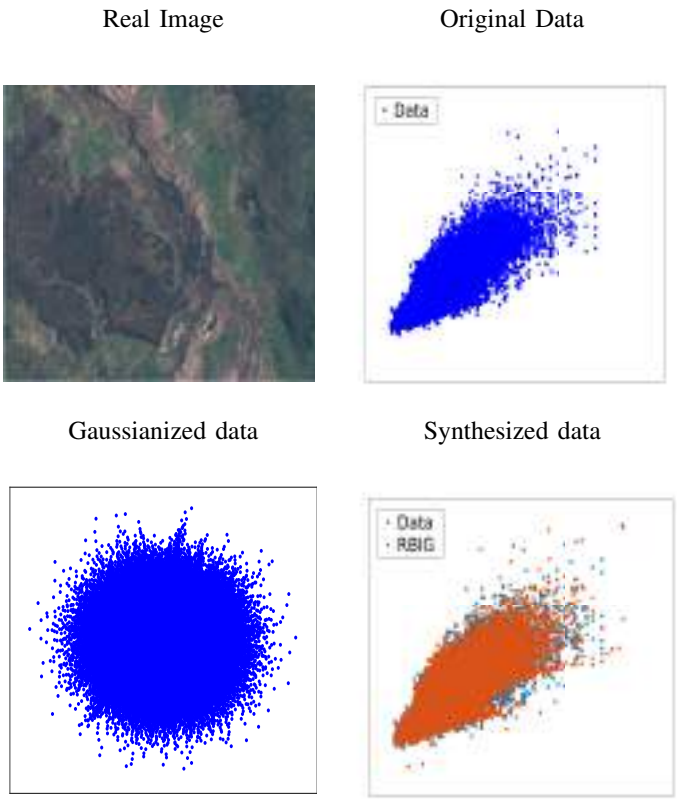


Fig. 1: Illustration of synthesized data using the RBIG methodology in a real Sentinel-2 image. Top-left: RGB composite of the original image. Top-right: representation of the first two bands of the original image. Bottom-left: first two dimensions of the Gaussianized data. Bottom-right: first two bands of the original image data (blue) and randomly generated data inverted using the learned Gaussianization transformation (orange).

to invert a kernel matrix, which has cubic cost with the number of points ℓ). Whereas some heuristics exist in the literature to fit the kernel parameters, in practice one only achieves the full potential of the KRX approach by fitting the parameters after cross-validation [10]. This requires having access to labeled data as anomalous versus non-anomalous classes, which is not a very realistic and not even practical setting. In this work, we approach the more useful and practical, yet more challenging, problem of unsupervised anomaly detection (i.e. no labeled data available), and therefore in our comparisons we will fit the kernel method parameter using the most successful (and sensible) heuristic to set the Gaussian kernel lengthscale σ as the average of all distances among $\mathbf{X}$.

As an alternative to linear measures of anomalousness like in RX, or nonlinear yet implicit feature transformations with parameters to tune like in KRX, we here propose a more straightforward approach to estimate the probability density function with RBIG (sec. II). This will give us a nonparametric parameter-free and efficient estimation of the data distribution. RBIG has optimal way of fitting the parameters of the distribution that do not require labeled data, and scales linearly with the data. By using RBIG to compute p_X, we obtain the method proposed in this work:

$$A_{\text{RBIG}}(\mathbf{x}) \propto \frac{1}{p_{\text{RBIG}}(\mathbf{x})}. \tag{6}$$

An important aspect to take into account is the intrinsic characteristics of the data used to estimate the density, which has implications in the quality of the estimation. When the distribution contains even a moderate number of anomalies, an accurate density estimate will cast anomalies as regular points, i.e. non-anomalous. This vastly depends on the flexibility of the class of models used. When the model is rigid like in the RX case, this is not a problem since it cannot be adapted to the anomalies. For the KRX one can control this effect by tuning the kernel lengthscale and the regularization term, but as explained before requires labeled data. This is an important aspect to take into account mostly in the anomaly detection scenario, where all data (included the anomalous samples) are used to estimate the density. Therefore we propose to use an hybrid model that combines the (too rigid) RX model with the (too flexible) RBIG model. The hybrid model first selects the data more likely not to be anomalous using RX and then uses this data to learn the Gaussianization transform with the RBIG model. This tries to avoid using anomalous data to train RBIG, which after all is intended to learn the background or pervasive data distribution. The number of data points selected as non-anomalous in the first step will define the trade-off between flexibility and rigidity.

B. RBIG for change detection

Change detection can be approached by setting thresholds on the change image (i.e. the difference between the two subsequent images for optical imagery or ratios in radar imagery) or from a purely density estimation standpoint. We will approach it from the latter angle using RBIG. This is certainly a more challenging approach, but has several associated advantages: 1) only the first image (or all previous images before the changed one) is considered to estimate the regular/background density; 2) there is no need to corregister images since the method operates in the geometric space defined by the image, not in the spatial domain; and 3) unlike a discriminative approach, a generative model like RBIG will allow us to derive useful descriptors of the image statistics, as well as to be refined as more images are acquired.

The idea to exploit RBIG for change detection is using data coming from the first image $\mathbf{X}_1$ only to estimate the probability model and then evaluating the probability (or change score, C) for each point in the second image $\mathbf{X}_2$, as follows:

$$C(\mathbf{x}_2) \propto \frac{1}{p_{X_1}(\mathbf{x}_2)}. \tag{7}$$

As for the anomaly detection case, we can use different models to estimate p_{X_1}. The most widely used is the Gaussian model. As in the previous section, when assuming a Gaussian distribution for the input data, the RX method can be used here too, i.e. $C_{\text{RX}}(\mathbf{y})$.

Likewise, kernel methods have been proposed to alleviate the strict assumption of Gaussian distribution [10] . While

different configurations were proposed in order to take into account only the anomalous changes, here we use the configuration designed for change detection. Following the idea in equation (7), the data of the first image ($\mathbf{X}_1$) is used to estimate the kernel and then the method is evaluated in the second image:

$$C_{\text{KRX}}(\mathbf{x}_2) \propto \frac{1}{p_{\text{K}(X_1)}(\mathbf{x}_2)}. \tag{8}$$

Equivalently, we can use RBIG to estimate the probability of the first image and evaluate the probability in the second one:

$$C_{\text{RBIG}}(\mathbf{x}_2) \propto \frac{1}{p_{\text{RBIG}(X_1)}(\mathbf{x}_2)}. \tag{9}$$

It is important to note that, in this case, the data used to estimate the probability density does not contain anomalies (changes in this setting) so the hybrid model is not needed here.

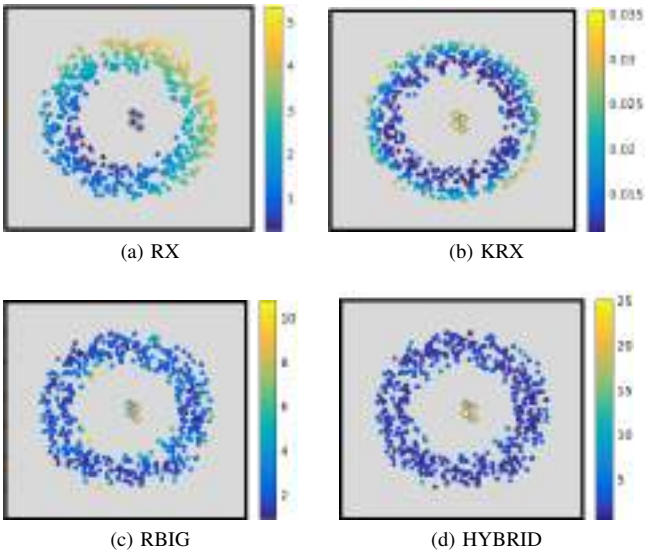
III. Experimental Results

This section analyzes the performance of the proposed RBIG method for anomaly and change detection. In order to assess the robustness we performed tests in both simulated and real scenes of varying dimensionality and sample size. We evaluate the detection power of the methods quantitatively through the Receiver Operating Characteristic (ROC) and Precision-Recall (PR) curves, along with the Area Under the Curve (AUC) scores. Besides, we provide examples of detection maps of each method to evaluate their quality by visual inspection.

We have performed three experiments. The first experiment is designed to illustrate the effect of the evaluated in an anomaly detection (AD) toy example. The second experiment deals with AD problem in different real scenarios: detection of air planes, latent fires, vehicles, and urbanization (roofs). The third experiment is related to evaluate the methods in change detection (CD) problems involvin floods, fires and droughts. Table I summarizes the different data sets used in the experiments. In order to ease the reproducibility, we provide MATLAB code implementations of the all methods. Moreover we made available a database with the labeled images used in the second and third experiments in https://isp.uv.es/RBIG4AD.html.

TABLE I: Image attributes used for the experiments of anomaly detection (AD) and change detection (CD).

Images	Sensor	Size	Bands	SR [m]
AD				
Cat-Island	AVIRIS	150×150	188	17.2
WTC	AVIRIS	200×200	224	1.7
Texas-Coast	AVIRIS	100×100	204	17.2
GulfPort	AVIRIS	100×100	191	3.4
CD				
Texas	Cross-Sensor	301×201	7	30
Argentina	Sentinel-2	1257×964	12	10-60
Chile	Landsat-8	201×251	12	10-60
Australia	Sentinel-2	1175×2031	12	10-60



(a) RX (b) KRX (c) RBIG (d) HYBRID

Fig. 2: Synthetic experiment to illustrate the methods performance when detecting anomalies. The color bar shows the intensity in terms of anomaly score from dark blue (less) to yellow (more). The image (a) correspond to RX detector, image (b) is the kernel version of RX, (c) represent the RBIG method and (d) showcase the hybrid model.

A. Experiment 1: Simulated Anomalies

The aim of this experiment is to illustrate the behavior of the proposed methods in challenging distributions exhibiting highly nonlinear feature relations. We designed a two-dimensional dataset where the non-anomalous data is in a circumference and the anomalous data in the middle. Figure 2 shows the performance of the different methods. The RX method assumption does not hold (the data is clearly non-Gaussian), hence it shows poor performance. The performance of KRX is better than RX but some false detections emerge in the outer circle, mainly related to the difficulty to select a reasonable kernel parameter. The direct application of RBIG easily identifies the anomalous points since they are far from the more dense (most probable) region. The proposed hybrid model further refines the detection since the density is estimated from pervasive data yielded by RX only.

B. Experiment 2: Anomaly Detection in Real Scenarios

We performed tests in four real examples. Table I summarizes relevant attributes of the datasets such as sensors, spatial and spectral resolution.

1) Data collection: We collected multispectral and hyperspectral images acquired by the AVIRIS and ROSIS-03 sensors. Figure 3 showcases the scenes used in the experiments. The AD scenarios consider anomalies related to a diversity of problems: airplane, latent fires, urbanization and vehicle detection [17]–[19].

The Cat-Island dataset corresponds to the airplane captured flying over the beach and it is considered a strange object

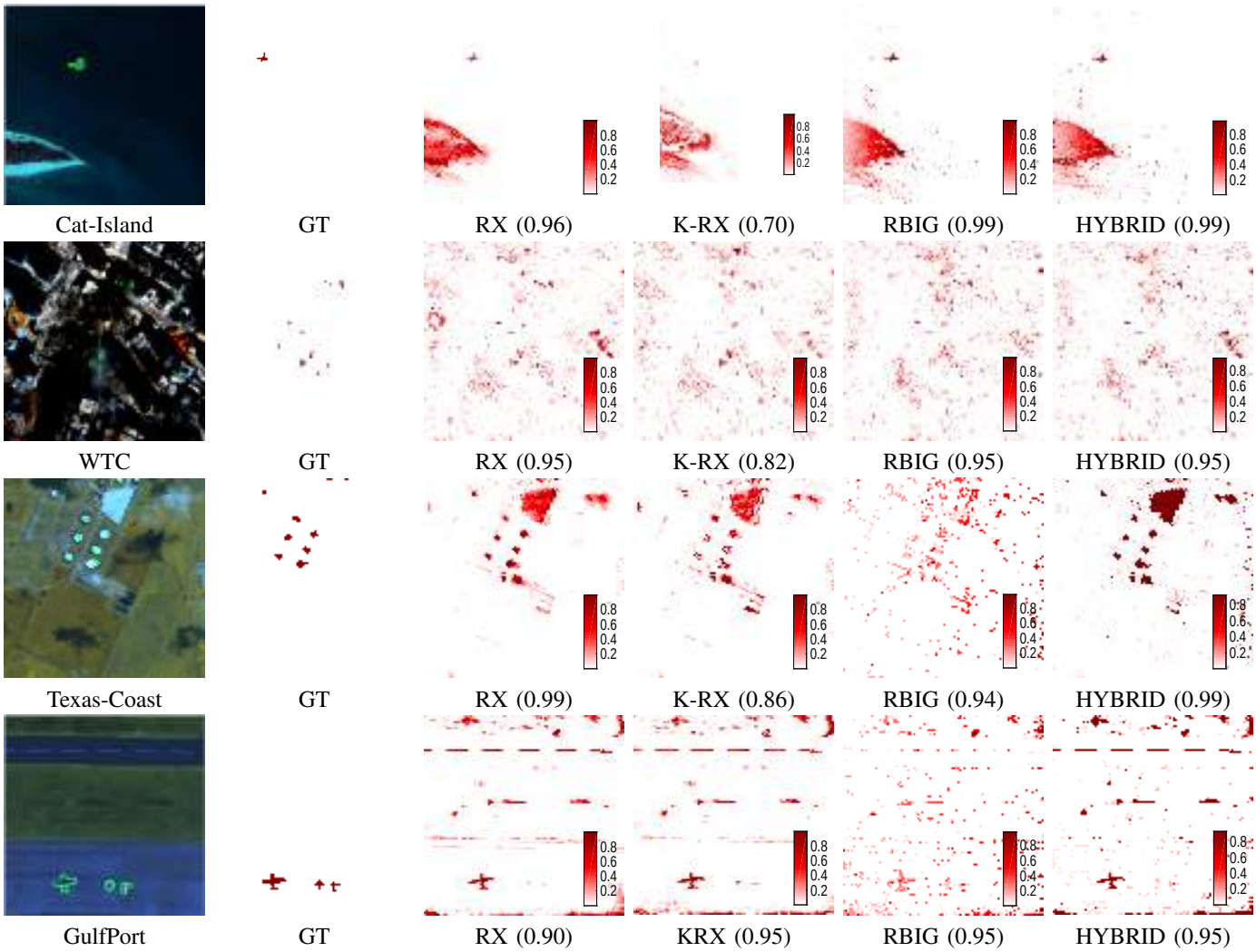


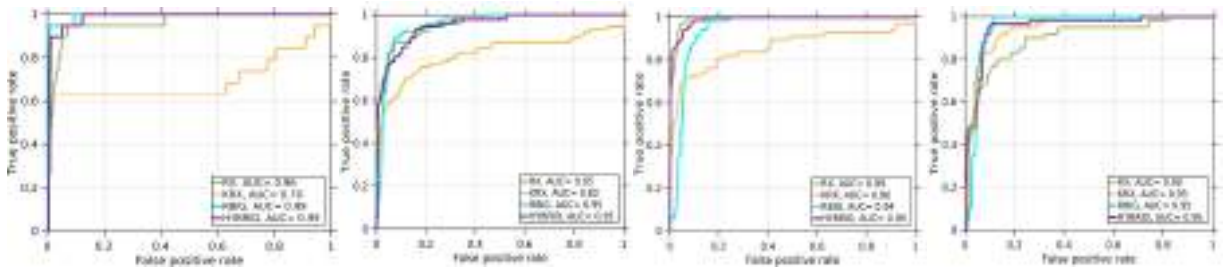
Fig. 3: Anomaly detection predictions in four images (one per row). First column: Cat-Island, World Trade Center (WTC), Texas Coast and Pavia original datasets with anomalies outlined in green. Second column: represent the reference maps of each image. From third column to the last column: activation maps and the AUC values (in parenthesis) for the RX, KRX, RBIG and the HYBRID models, respectively.

when compared to the rest of the image (a white spot in the middle of a beach) and the percentage of anomalies represent the 0.09% of the scene. The World Trade Center (WTC) image was collected by the Airborne Visible Infra-Red Imaging Spectrometer (AVIRIS) over the WTC area in New York on 16 September 2001 (after the collapse of the towers in NY). The data set covered the hot spots corresponding to latent fires at the WTC, which can be considered as anomalies and it represent the 0.23% of the scene. In the Texas Coast dataset, the anomalies represent the 0.67% of the scene and the image contains roofs built on a wooded site and bright spots that reflect light which can be considered an anomaly. The GulfPort dataset correspond to a battery of airplanes taxied on the runway and the pecentage of anomalies represent the 0.60% of the scene.

2) Numerical and Visual Comparison: It is important to take into consideration that KRX requires the selection of some hyperparameters, being the kernel parameter the most critical one. In order to perform a fair comparison while staying in an unsupervised learning setting, we use the standard RBF kernel function, $k(\mathbf{a}, \mathbf{b}) = \exp(-\|\mathbf{a} - \mathbf{b}\|^2/(2\sigma^2))$ and set the lengthscale parameter σ to the median distance between all examples. A visual comparison of the results in terms of activation maps for all methods is given in Fig. 3. They display the predictions given to each sample. The prediction maps show a binary representation between change and non-change

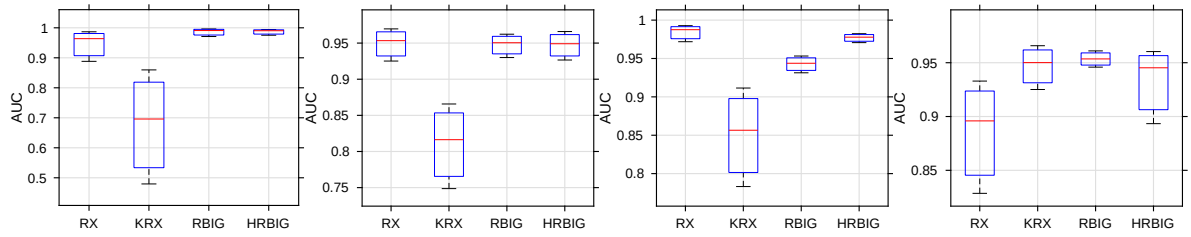
TABLE II: AUC results for Anomaly Detection images. The value for the best method for each image is in bold.

METHODS	RX	K-RX	RBIG	HYBRID
Cat-Island	*0.96*	*0.70*	**0.99**	**0.99**
WTC	**0.95**	*0.82*	**0.95**	**0.95**
Texas-Coast	**0.99**	*0.86*	*0.94*	**0.99**
GulfPort	*0.90*	**0.95**	**0.95**	**0.95**



(a) Cat-Island (b) WTC (c) Texas Coast (d) GulfPort

Fig. 4: Anomaly detection ROC curves in linear scale for all scenes. Numbers in legend display the AUC values for each method.



(a) Cat-Island (b) WTC (c) Texas Coast (d) GulfPort

Fig. 5: Anomaly detection results of the bootstrap experiment for 1000 experiments. AUC values and standard deviation for each method are shown as boxplot, red line represent the median value, the blue box contains 95% of the values, black lines represent the maximum and minimum values.

samples obtained from the model subject to a threshold. Results in all scenes demonstrate that (1) RX is a competitive method for detection, (2) KRX struggles to obtain reasonable results mainly due to the problem of hyperparameter tuning, (3) RBIG alone excels in all cases, while the hybrid approach (i.e. RX followed by RBIG) refines the results and yields clearer activation maps with sharper spatial detections.

Additionally, for a quantitative assessment of the results, it is customary to provide the ROC curves and to derive scores like the AUC from it. Figure 4 shows the ROC curves and Table II summarizes all AUC values for all images and methods. For each experiment, we performed 1000 runs for testing the significance of the methods based on the ROC profiles. The results are shown in Figure 5. Although the RBIG model achieves good results, RX model is able to compete and achieve results as good as RBIG for some images. The HYBRID model is able to keep the properties of the above mentioned models obtaining results equal or better than any other method. While KRX obtains a reasonable performance in some images, it clearly fails in some situations like the Cat-Island image. The low standard deviations show that all methods but the KRX are clearly robust with a little bit bigger standard deviation for the RX method in most cases.

C. Experiment 3: Real and Natural Changes

This section reports an experiment to analyze the performance of the proposed methods in change detection problems. The database is composed of different scenes with natural changes, whose characteristics are summarized in Table I.

1) Data collection: We collected pairs of multispectral images in such a way that they coincide at the same spatial resolution but at different acquisition time, the images are co-registered. We selected the images in such a way that an anomalous change happened between the two acquisition times. We manually labeled all the images finding the changed pixels. Labeling considered avoiding shadows, changes in lighting and natural changes in vegetation which could compromise results evaluation. All images contain changes of different nature, which allows us to analyze and study how the algorithms perform in heterogeneous realistic scenarios. The Texas wildfire dataset is composed by a set of four images acquired by different sensors over Bastrop County, Texas (USA), and is composed by a Landsat 5 TM as the pre-event image and a Landsat 5 TM plus an EO-1 ALI and a Landsat 8 as post-event images. This phenomenon is considered the most destructive wildland-urban interface wildfire in Texas history and the interest region represent the 19.54%. The Argentina image represents an area burned between the months of July and August 2016 due to the high temperatures in these crop areas, the change region representing the 7.5% of the whole scene. The Chile dataset represents the Aculeo lake in central part of this country, which has now dried up completely. These images contrast the lake in 2014, when it still contained substantial water, and 2019, when it consisted of dried mud and green vegetation. Scientists attribute the lake's decline to an unusual decade-long drought, coupled with increased water consumption from a growing population, and the changed region represents a relevant 10.81% of the whole

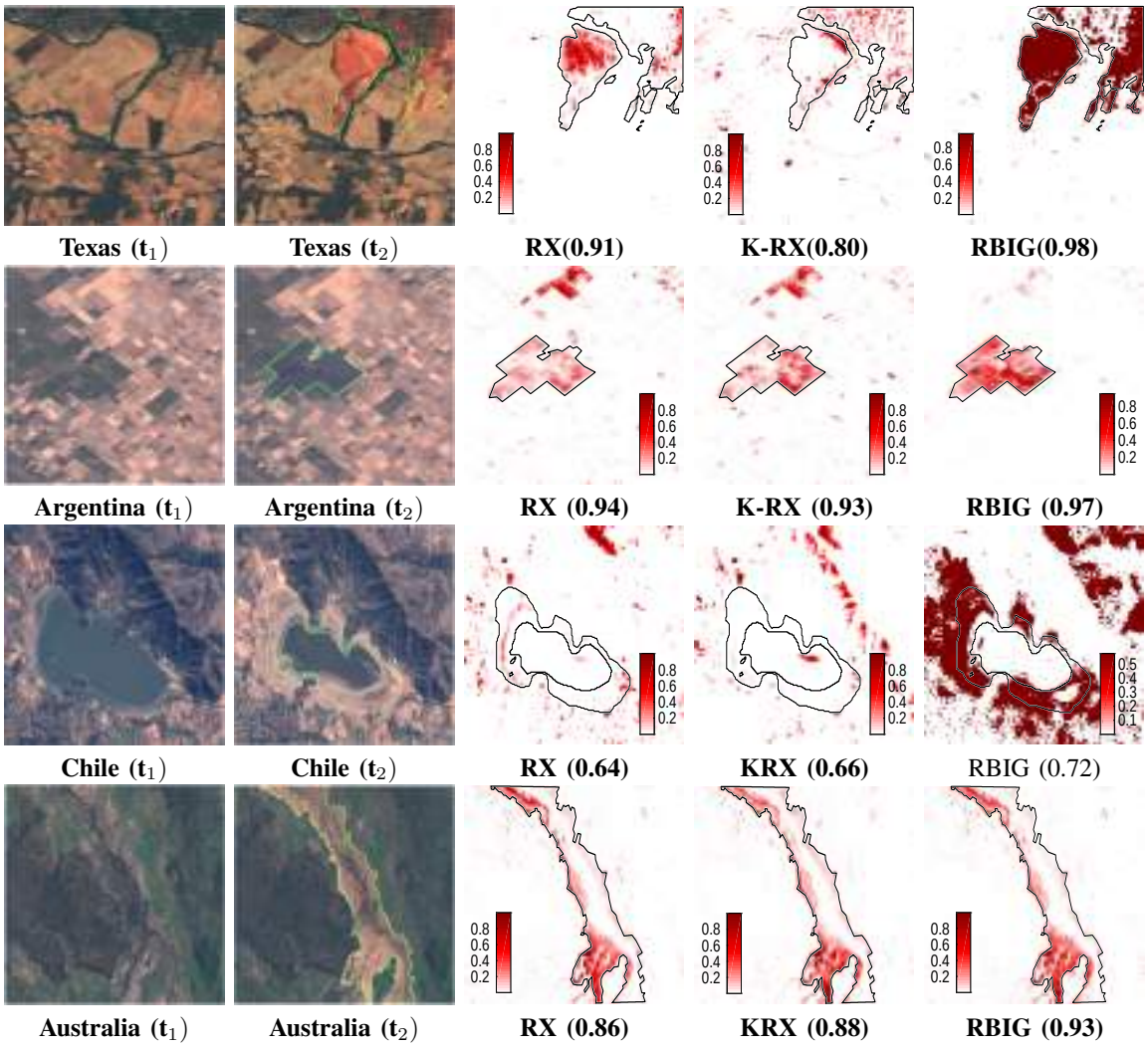


Fig. 6: Change detection results for different images. First two columns show the images before and after the change, with the changed region highlighted in green. Columns three to five show the prediction maps for the different methods, the amount of change detected in each pixel is colored from white (less) to red (more). AUC values are given in parenthesis. The changed region is outlined in black to facilitate the visual inspection.

scene. The last dataset labeled as Australia shows the natural floods caused by Cyclone Debbie in Australia 2017. Storm damage resulted from both the high winds associated with the cyclone, and the very heavy rain that produced major riverine floods. The change samples represent an important portion of the scene, the 17.35% of pixels affected. Since our RBIG approach only takes the time t_1 image, these big changes do not have a critical impact on method's performance.

2) Numerical and Visual Comparison: Figure 6 shows the RGB composites of the pairs of images, the corresponding reference map and activation maps obtained. RBIG obtains clearly better results than the other methods in all cases; very good performance in three out of the four scenarios and a clear advantage in the most difficult one (Chile image). When dealing with highly skewed datasets, PR curves give a more informative picture of an algorithm's performance compared to ROC. Figure 7 shows both the ROC and the PR curves results for all methods and all the images. In all cases RBIG outperforms the other methods largely, thus suggesting the suitability of adopting a more direct approach of density estimation in the change detection problems too. A summary of the AUC values of all methods and scenarios is shown in Table III. The RBIG approach is to able to estimate the change samples with a high accuracy overtaking in 7%, 3%, 6% and 5% respectively with respect the second best method.

TABLE III: AUC results for Change Detection images. The best value for each image are in bold

METHODS	RX	K-RX	RBIG
Texas	*0.91*	*0.80*	**0.98**
Argentina	*0.94*	*0.93*	**0.97**
Chile	*0.64*	*0.66*	**0.72**
Australia	*0.86*	*0.88*	**0.93**

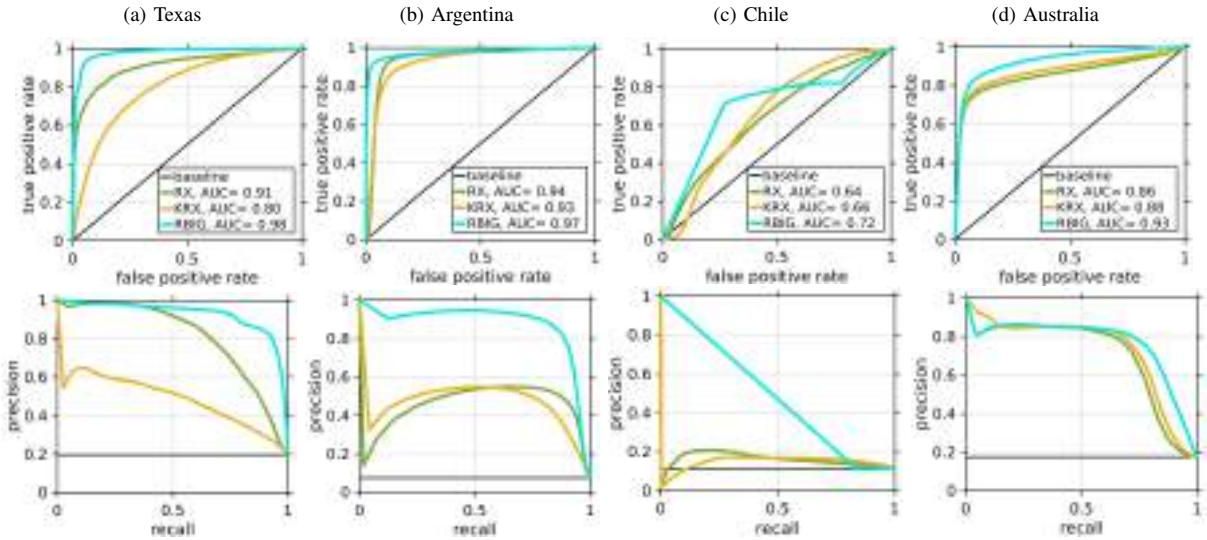


Fig. 7: ROC (top row) and Precision-Recall (bottom row) curves for change detection problems.

IV. Conclusions

We introduced a novel detector based on multivariate Gaussianization. The methodology copes with anomaly and change detection problems in remote sensing image processing, and meets all requirements of the problems: is an unsupervised method with no parameters to fit, it can deal with large amount of data, and it is more accurate to competing approaches. The model assumption is based on detecting anomalies by estimating probabilities of pixels. The proposed method excelled quantitatively (AUC, ROC and PR curves) and qualitative based on visual inspection over the rest of the implementations, in both anomaly and change detection. The evaluation considered a wide range of remote sensing images, in a diversity of problems, dimensionality and number of examples. We also suggested a hybrid approach where the Gaussianization method is applied after a regular anomaly detector: this facilitates the density estimation and improves the results notably. Future work will consider exploiting the information-theoretic properties of RBIG [20] which opens alternatives to identify changes in image time series.

References

[1] G. Camps-Valls, D. Tuia, L. Bruzzone, and J. A. Benediktsson, "Advances in hyperspectral image classification: earth monitoring with statistical learning methods," *IEEE Signal Processing Magazine*, vol. 31, no. 1, pp. 45–54, 2014.

[2] Antonio Plaza, Jon Atli Benediktsson, and Joseph W. Boardman, "Recent advances in techniques for hyperspectral image processing," *Remote Sensing of Environment*, vol. 113, pp. S110 – S122, 2009.

[3] G. Camps-Valls, Devis Tuia, Luis Gómez-Chova, Sandra Jiménez, and Jesús Malo, *Remote sensing image processing*, Morgan & Claypool Publishers, 1st edition, 2011.

[4] D. W. J. Stein, S. G. Beaven, L. E. Hoff, E. M. Winter, A. P. Schaum, and A. D. Stocker, "Anomaly detection from hyperspectral imagery," *IEEE Signal Processing Magazine*, vol. 19, no. 1, pp. 58–69, Jan 2002.

[5] S. Matteoli, M. Diani, and G. Corsini, "A tutorial overview of anomaly detection in hyperspectral images," *IEEE Aerospace and Electronic Systems Magazine*, vol. 25, no. 7, pp. 5–28, 2010.

[6] M. Ben Salem, K. Saheb Ettabaa, and M. S. Bouhlel, "An adaptive spatial and spectral neighborhood for the rx anomaly detector," in *IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium*, 2018, pp. 8484–8487.

[7] Y. E. Sahın, S. Arisoy, and K. Kayabol, "Anomaly detection with bayesian gauss background model in hyperspectral images," in *2018 26th Signal Processing and Communications Applications Conference (SIU)*, 2018, pp. 1–4.

[8] I. S. Reed and X. Yu, "Adaptive multiple-band CFAR detection of an optical pattern with unknown spectral distribution," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 38, pp. 1760–1770, 1990.

[9] Heesung Kwon and N. M. Nasrabadi, "Hyperspectral anomaly detection using kernel RX-algorithm," in *2004 International Conference on Image Processing, 2004. ICIP '04.*, 2004, vol. 5, pp. 3331–3334 Vol. 5.

[10] J. A. Padron-Hidalgo, A. Perez-Suay, F. Nar, and G. Camps-Valls, "Efficient Nonlinear RX Anomaly Detectors," *IEEE Geoscience and Remote Sensing Letters*, pp. 1–5, 2020.

[11] V. Laparra, G. Camps-Valls, and J. Malo, "Iterative Gaussianization: From ICA to Random Rotations," *IEEE Transactions on Neural Networks*, vol. 22, pp. 537–549, 2011.

[12] D. Lu, P. Mausel, E. Brondízio, and E. Moran, "Change detection techniques," *International Journal of Remote Sensing*, vol. 25, no. 12, pp. 2365–2401, 2004.

[13] Jerome H. Friedman, "Exploratory Projection Pursuit," 1987, vol. 82, pp. 249–266.

[14] Scott Saobing Chen and Ramesh A. Gopinath, "Gaussianization," in *NIPS*, 2000.

[15] G. Camps-Valls and L. Bruzzone, *Kernel methods for Remote Sensing Data Analysis*, Wiley & Sons, UK, Dec 2009.

[16] J.L. Rojo-Álvarez, M. Martínez-Ramón, J. Muñoz-Marí, and G. Camps-Valls, *Digital Signal Processing with Kernel Methods*, Wiley & Sons, UK, Apr 2018.

[17] Qiandong Guo, Ruiliang Pu, and Jun Cheng, "Anomaly detection from hyperspectral remote sensing imagery," *Geosciences*, vol. 6, no. 4, 2016.

[18] X. Kang, X. Zhang, S. Li, K. Li, J. Li, and J. A. Benediktsson, "Hyperspectral anomaly detection with attribute and edge-preserving filters," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 10, pp. 5600–5611, 2017.

[19] J. A. Padrón-Hidalgo, V. Laparra, N. Longbotham, and G. Camps-Valls, "Kernel anomalous change detection for remote sensing imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 10, pp. 7743–7755, Oct 2019.

[20] V. Laparra, J. E. Johnson, G. Camps-Valls, R. Santos-Rodriguez, and J. Malo, "Information theory measures via multidimensional gaussianization," 2020, Available at arxiv:2010.03807.

Bibliography

Ahlberg, J., & Renhorn., I. (2004). Multi- and hyperspectral target and anomaly detection. *Tech Report FOI-R-1526-SE, Swedish Defence Research Agency*, .

Alaoui, A., & Mahoney, M. W. (2015). Fast Randomized Kernel Ridge Regression with Statistical Guarantees. In *Advances in Neural Information Processing Systems* (pp. 775–783). volume 28.

Arenas-Garcia, J., Petersen, K. B., Camps-Valls, G., & Hansen, L. K. (2013). Kernel multivariate analysis framework for supervised subspace learning: A tutorial on linear and kernel multivariate methods. *IEEE Signal Processing Magazine*, *30(4)*, 16–29.

Bae, W., S., H., & C., K. (2008). Influence diagnostics in the varying coefficient model with longitudinal data. Computational Statistics, . *23*, 185–196.

Benz, U. C., Hofmann, P., Willhauck, G., Lingenfelder, I., & Heynen, M. (2004). Multi-resolution, object-oriented fuzzy analysis of remote sensing data for GIS-ready information. *ISPRS Journal of photogrammetry and remote sensing*, *58*, 239–258.

Bovolo, F. (2009). A Multilevel Parcel-Based Approach to Change Detection in Very High Resolution Multitemporal Images. *IEEE Geoscience and Remote Sensing Letters.*, *6(1)*, 33–37.

Bovolo, F., & Bruzzone, L. (2007). A Split-Based Approach to Unsupervised Change Detection in Large-Size Multitemporal Images: Application to Tsunami-Damage Assessment. *IEEE Transactions on Geoscience and Remote Sensing*, *45(6)*, 1658–1670.

Bovolo, F., Bruzzone, L., & Marchesi, S. (2009). Analysis and Adaptive Estimation of the Registration Noise Distribution in Multitemporal VHR Images. *IEEE Transactions on Geoscience and Remote Sensing*, *47(8)*, 2658–2671.

Cambanis, S., Huang, S., & Simons, G. (1981). On the theory of elliptically contoured distributions. *J. Multiv. Anal.*, *11*, 368–385.

Campos-Taberner, M., García-Haro, F. J., Camps-Valls, G., Grau-Muedra, G., Nutini, F., Crema, A., & Boschetti, M. (2016). Multitemporal and multiresolution leaf area index retrieval for operational local rice crop monitoring. *Remote Sensing of Environment*, *187*, 102 – 118.

Camps-Valls, B. L. M., G., Kottas, A., Taddy, M., & Ganapol, B. D. (2009). *Kernel Methods for Remote Sensing Data Analysis.*.

Camps-Valls, G. (2016). Kernel spectral angle mapper. *IEE Electronics Letters*, *52*, 1218–1220. URL: http://digital-library.theiet.org/content/journals/10.1049/el.2016.0661. doi:http://dx.doi.org/10.1049/el.2016.0661.

Camps-Valls, G., & Bruzzone, L. (2009.). Kernel Methods for Remote Sensing Data Analysis. *Wiley & Sons, UK*, .

Camps-Valls, G., Mooij, J. M., & Schölkopf, B. (2010). Remote Sensing Feature Selection by Kernel Dependence Measures. *IEEE Geoscience and Remote Sensing Letter*, *7*, 587–591.

Canty, M. J., & Nielsen, A. A. (2008). Automatic radiometric normalization of multitemporal satellite imagery with the iteratively re-weighted MAD transformation. *Remote Sensing of Environment*, *112(3)*, 1025–1036.

Cao, C., Yu, J., Zhou, C., Hu, K., Xiao, F., & Gao, X. (2019). Hyperspectral image denoising via subspace-based nonlocal low-rank and sparse factorization. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *12(3)*, 973–988.

Celik, T. (2009). Multiscale Change Detection in Multitemporal Satellite Images. *IEEE Geoscience and Remote Sensing Letters*, *6(4)*, 820–824.

Chang, C., & Chiang, S. (2002). Anomaly detection and classification for hyperspectral imagery. *IEEE Trans. Geosci. Rem. Sens.*, *40*, 1314–1325.

Chen, J., Chen, X., Cui, X., & Chen, J. (2010). Change Vector Analysis in Posterior Probability Space: A New Method for Land Cover Change Detection. *IEEE Geoscience and Remote Sensing Letters*, *8(2)*, 317–321.

Chen, S. S., & Gopinath, R. A. (2000). Gaussianization. In *NIPS* (pp. 423–429). volume 13.

Choongrak, K., & Barry, E. S. (1996). Reference values for Cook's distance. *Communications in Statistics - Simulation and Computation*, *25*, 691–708.

Cook, R. D. (1977). Detection of influential observation in linear regression. *Technometrics*, *19(1)*, 15 – 18.

Cook, R. D., & Weisberg, S. (1980). Characterizations of an Empirical Influence Function for Detecting Influential Cases in Regression. *Technometrics*, *22*, 495–508. doi:10.1080/00401706.1980.10486199.

Coppin, P., Jonckheere, I., Nackaerts, K., & B., M. (2004). Digital change detection methods in ecosystem monitoring: A review. *International Journal of Remote Sensing*, *25(9)*, 1565–1596.

D. Manolakis, D. Z., L. G. Jairam, & Rossacci, M. (2007). Statistical models for LWIR hyperspectral backgrounds and their applications in chemical agent detection. (p. 656525). volume 6565. doi:10.1117/12.718378.

Dalla Mura, M., Benediktsson, J. A., Bovolo, F., & Bruzzone, L. (2008). An Unsupervised Technique Based on Morphological Filters for Change Detection in Very High Resolution Images. *IEEE Geoscience and Remote Sensing Letters*, *5(3)*, 433–437.

Danson, F. M., & Plummer, S. E. (1995). *Advances in Environmental Remote Sensing*. New York: John Wiley & Sons.

Ding, M., Tian, Z., Jin, Z., Xu, M., & Cao, C. (2010). Registration Using Robust Kernel Principal Component for Object-Based Change Detection. *IEEE Geoscience and Remote Sensing Letters*, *7(4)*, 761–765. doi:10.1109/LGRS.2010.2047241.

Donlon, C., Berruti, B., Buongiorno, A., Ferreira, M.-H., Féménias, P., Frerick, J., Goryl, P., Klein, U., Laur, H., Mavrocordatos, C., Nieke, J., Rebhan, H., Seitz, B., Stroede, J., & Sciarra, R. (2012.). The global monitoring for environment and security (GMES) sentinel-3 mission. *Remote Sensing Environment.*, *120*, 37–57.

Drusch, M., Bello, U. D., Carlier, S., Colin, O., Fernandez, V., Gascon, F., Hoersch, B., Isola, C., Laberinti, P., Martimort, P., Meygret, A., Spoto, F., Sy, O., Marchese, F., & Bargellini, P. (2012.). Sentinel-2: Esa's optical High-Resolution Mission for GMES Operational Services. *Remote Sensing Environment.*, *120*, 25–86.

Eubank, R. L. (1985). Diagnostics for smoothing splines. *Journal of the Royal Statistical Society. Series B (Methodological)*, *47*, 332–341.

Fine, S., & Scheinberg, K. (2001). Efficient SVM training using low-rank kernel representations. *Journal of Machine Learning Research*, *2*, 243–264.

Friedman, J. H. (1987). Exploratory Projection Pursuit. (pp. 249–266). volume 82.

Fung, W.-K., Zhu, Z.-Y., Wei, B.-C., & He, X. (2002). Influence diagnostics and outlier tests for semiparametric mixed models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *64(3)*, 565–579.

Ghosh, S., Bruzzone, L., Patra, S., Bovolo, F., & Ghosh, A. (2007). A Context-Sensitive Technique for Unsupervised Change Detection Based on Hopfield-Type Neural Networks. *IEEE Transactions on Geoscience and Remote Sensing*, *45(3)*, 778–789.

Gómez-Chova, L., Muñoz Marí, J., Laparra, V., Malo-López, J., & Camps-Valls, G. (2011). A review of kernel methods in remote sensing data analysis. In S. Prasad, L. M. Bruce, & J. Chanussot (Eds.), *Optical Remote Sensing – Advances in Signal Processing and Exploitation Techniques* (pp. 171–206). Springer Berlin Heidelberg volume 3 of *Augmented Vision and Reality*. URL: http://dx.doi.org/10.1007/978-3-642-14212-3_10. doi:10.1007/978-3-642-14212-3_10.

Gómez-Chova, L., Nielsen, A. A., & Camps-Valls, G. (2011). Explicit signal to noise ratio in reproducing kernel Hilbert spaces. In *2011 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)* (pp. 3570–3573). IEEE.

Green, A. A., Berman, M., Switzer, P., & Craig, M. D. (1998). A transformation for ordering multispectral data in terms of image quality with implications for noise removal. *IEEE Trans. Geosc. Rem. Sens.*, *26*, 65–74.

Gretton, A., Borgwardt, K., Rasch, M., Schölkopf, B., & Smola., A. (2012). A Kernel Two-Sample Test. *Journal of Machine Learning Research.*, *13*, 723–773.

Guo, Q., Pu, R., & Cheng, J. (2016). Anomaly detection from hyperspectral remote sensing imagery. *Geosciences*, *6(4)*, 56. doi:10.3390/geosciences6040056.

Heesung Kwon, & Nasrabadi, N. M. (2004). Hyperspectral anomaly detection using kernel RX-algorithm. In *2004 International Conference on Image Processing, 2004. ICIP '04.* (pp. 3331–3334). volume 5.

Hoerl, A. E., & Kennard, R. W. (1970). Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, *12*, 55–67. doi:10.1080/00401706.1970.10488634.

Im, J., Jensen, J. R., & Hodgson, M. E. (2008). Optimizing the binary discriminant function in change detection applications. *Remote Sensing of Environment.*, *112(6)*, 2761 – 2776.

IPCC (2012). Managing the Risks of Extreme Events and Disasters to Advance Climate Change Adaptation. Cambridge University Press, Cambridge, UK, and New York, NY, USA, . (p. 582).

Justice, C. O., Vermote, E., Townshend, J. R., Defries, R., Roy, D. P., Hall, D. K., Salomonson, V. V., Privette, J. L., Riggs, G., Strahler, A. et al. (1998). The moderate resolution imaging spectroradiometer (MODIS): Land remote sensing for global change research. *IEEE Transactions on Geoscience and Remote Sensing*, *36(4)*, 1228–1249.

Kang, X., Zhang, X., Li, S., Li, K., Li, J., & Benediktsson, J. A. (2017). Hyperspectral anomaly detection with attribute and edge-preserving filters. *IEEE Transactions on Geoscience and Remote Sensing*, *55(10)*, 5600–5611. doi:10.1109/TGRS.2017.2710145.

Keshava, N. (2004). Distance metrics and band selection in hyperspectral processing with applications to material identification and spectral libraries. *IEEE Transaction on Geoscience and Remote Sensing*, *42(7)*, 1552–1565.

Kim, C., Y., L., & U., P. B. (2001). Cook's distance in local polynomial regression. *Statistics & Probability Letters*, *54(1)*, 33 – 40.

Kraft, S., Del Bello, U., Bouvet, M., Drusch, M., & Moreno, J. (2012.). FLEX: ESA's earth explorer 8 candidate mission. *IEEE International Geoscience and Remote Sensing Symposium*, (pp. 7125–7128).

Kumar, N. K., & Schneider, J. (2017). Literature survey on low rank approximation of matrices. *Linear and Multilinear Algebra*, *65(11)*, 2212–2244.

Kwon, H., Der, S. Z., & Nasrabadi, N. M. (2003). Adaptive anomaly detection using subspace separation for hyperspectral imagery. *Optical Engineering*, *42*, 3342 – 3351.

Kwon, H., & Nasrabadi, N. M. (2005). Kernel RX - algorithm: a nonlinear anomaly detector for hyperspectral imagery. *IEEE Transaction on Geoscience and Remote Sensing*, *43(2)*, 388–397.

Labate, D., Ceccherini, M., Cisbani, A., Cosmo, V. D., Galeazzi, C., Giunti, L., Melozzi, M., Pieraccini, S., & Stagi, M. (2009.). The PRISMA payload optomechanical design, a

high performance instrument for a new hyperspectral mission. *Acta Astronautica*, *65*, 1429–1436.

Laparra, V., Camps-Valls, G., & Malo, J. (2011). Iterative gaussianization: From ica to random rotations. *IEEE Transactions on Neural Networks*, *22*, 537–549.

Liang, S. (2004). *Quantitative Remote Sensing of Land Surfaces*. New York: John Wiley & Sons.

Lillesand, T., Kiefer, R. W., & Chipman, J. (2014). *Remote sensing and image interpretation*. John Wiley & Sons.

Liu, J., Gong, M., Qin, K., & Zhang, P. (2018). A Deep Convolutional Coupling Network for Change Detection Based on Heterogeneous Optical and Radar Images. *IEEE Transactions on Neural Networks and Learning Systems.*, *27(3)*, 545 – 559.

Longbotham, N., & Camps-Valls, G. (2014). A family of kernel anomaly change detectors. In *2014 6th Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS)* (pp. 1–4). doi:10.1109/WHISPERS.2014.8077648.

Longbotham, N., & Camps-Valls, G. (2014). A family of kernel anomaly change detectors,. *Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS)*, (pp. 1–4).

Longbotham, N., Pacifici, F., Baugh, B., & Camps-Valls, G. (2014). Prelaunch Assessment of WorldView-3 Information Content," in "6th Work. Hyperspectral Image Signal Process. *Evolution Remote Sensing. WHISPERS," (Lausanne, Switzerland, 2014) 24-27 June, 2014, Lausanne, Switzerland 2014*, .

Lu, D., Mausel, P., Brondrizio, E., & Moran., E. (2004). Change detection techniques. *International Journal of Remote Sensing*, *25(12)*, 2365–2401.

Lu, X., Hoff, L., Reed, I., Chen, M., & Stotts, L. (1997). Automatic target detection and recognition in multiband imagery: a unified ML detection and estimation approach. *IEEE Transactions on Image Processing*, *6(1)*, 143–156.

Lyu, S., & Simoncelli, E. P. (2009). Nonlinear extraction of 'independent components' of natural images using radial Gaussianization. *Neural Computation*, *21*, 1485–1519. doi:10.1162/neco.2009.04-08-773.

Malenovský, Z., Rott, H., Cihlar, J., Schaepman, M. E., García-Santos, G., Fernandes, R., & Berger, M. (2012). Sentinels for science: Potential of Sentinel-1, -2, and -3 missions for scientific observations of ocean, cryosphere, and land. *Remote Sensing of Environment*, *120*, 91 – 101.

Malila, W. A. (1980). Change Vector Analysis: An Approach for Detecting Forest Change with Landsat. In *IEEE Proceedings of Annual Symposium on Machine Processing of Remotely Sensing Data* (pp. 326–335).

Manolakis, D., Marden, D., & Shaw, G. A. (2003). Hyperspectral image processing for automatic target detection applications. *Lincoln Laboratory Journal*, *14(1)*, 79–116.

Manolakis, D., & Shaw, G. (2002). Detection algorithms for hyperspectral imaging applications. *IEEE Signal Processing Magazine*, *19(1)*, 29–43.

Mas, J.-F. (1999). Monitoring land-cover changes: A comparison of change detection techniques. *International journal of remote sensing*, *20(1)*, 139–152.

Matteoli, S., Diani, M., & Corsini, G. (2010). A tutorial overview of anomaly detection in hyperspectral images. *IEEE Aerospace and Electronic Systems Magazine*, *25(7)*, 5–28. doi:10.1109/MAES.2010.5546306.

Mayer, R., Bucholtz, F., & Scribner, D. (2003). Object detection by using "whitening/dewhitening" to transform target signatures in multitemporal hyperspectral and multispectral imagery. *IEEE Transactions on Geoscience and Remote Sensing*, *41(5)*, 1136–1142.

McCurdy, S. (2018). Ridge regression and provable deterministic ridge leverage score sampling. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems 31* (pp. 2463–2472).

Morales-Álvarez, P., Pérez-Suay, A., Molina, R., & Camps-Valls, G. (2018). Remote sensing image classification with large-scale Gaussian processes. *IEEE Transaction on Geoscience and Remote Sensing*, *56(2)*, 1103–1114.

Mott, H. (2007). *Remote Sensing with Polarimetric Radar*. New York: John Wiley & Sons.

Muñoz-Marí, J., Bovolo, F., Gómez-Chova, L., Bruzzone, L., & Camps-Valls, G. (2010). Semisupervised one-class support vector machines for classification of remote sensing data. *IEEE Transaction and Geoscience Remote Sensing.*, *48(8)*, 3188–3197.

Nar, F., Pérez-Suay, A., Padrón-Hidalgo, J. A., & Camps-Valls, G. (2018). Randomized RX for Target Detection. In *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)* (pp. 4237–4240). doi:10.1109/IGARSS.2018.8517961.

Nielsen, A. A. (2006). The Regularized Iteratively Reweighted MAD Method for Change Detection in Multi- and Hyperspectral Data. *IEEE Transactions on Image Processing*, *16(2)*, 463–478.

Nielsen, A. A. (2011). Kernel Maximum Autocorrelation Factor and Minimum Noise Fraction Transformations. *IEEE Transactions on Image Processing*, *20(3)*, 612–624.

Nielsen, A. A., Conradsen, K., & Simpson, J. J. (1998). Multivariate Alteration Detection MAD and MAF Postprocessing in Multispectral, Bitemporal Image Data: New Approaches to Change Detection Studies. *Remote Sensing of Environment*, *64(1)*, 1–19.

Ouyang, W., Zeng, X., Wang, X., Qiu, S., Luo, P., Tian, Y., Li, H., Yang, S., Wang, Z., Li, H., Wang, K., Yan, J., Loy, C., & Tang, X. (2017). Deepid-net: Object detection with deformable part based convolutional neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *39(7)*, 1320–1334.

Pacifici, F., Chini, M., Bignami, C., Stramondo, S., & Emery, W. (2010). Automatic damage detection Using pulse-coupled neural networks for the 2009 Italian earthquake. In *2010 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)* (pp. 1996–1999).

Pacifici, F., Del Frate, F., & Emery, W. (2009). Pulse Coupled Neural Networks for detecting urban areas changes at very high resolutions. In *2009 Joint Urban Remote Sensing Event* (pp. 1–7).

Padron-Hidalgo, J. A., Perez-Suay, A., Nar, F., & Camps-Valls, G. (2021). Efficient Nonlinear RX Anomaly Detectors. *IEEE Geoscience and Remote Sensing Letters*, *18(2)*, 231–235.

Padrón-Hidalgo, J. A., Laparra, V., Longbotham, N., & Camps-Valls, G. (2019). Kernel anomalous change detection for remote sensing imagery. *IEEE Transactions on Geoscience and Remote Sensing*, *57*, 7743–7755. doi:10.1109/TGRS.2019.2916212.

Pasquier, H., & Verheyden, A. (1998). The Vegetation Programming Center, a new subsystem in the Spot4 ground segment. In *DASIA 98-Data Systems in Aerospace* (p. 367). volume 422.

Radke, R. J., Andra, S., Al-Kofahi, O., & Roysam, B. (2005). Image change detection algorithms: a systematic survey. *IEEE Transactions on Image Processing*, *14(3)*, 294–307.

Rahimi, A., & Recht, B. (2007). Random Features for Large-Scale Kernel Machines. In *NIPS* (p. 5). Citeseer volume 3(4).

Reed, I. S., & Yu, X. (1990a). Adaptive multiple-band CFAR detection of an optical pattern with unknown spectral distribution. *IEEE Transaction on Acoustics, Speech and Signal Processing*, *38*, 1760–1770. doi:10.1109/29.60107.

Reed, I. S., & Yu, X. (1990b). Adaptive multiple-band cfar detection of an optical pattern with unknown spectral distribution. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *38(10)*, 1760–1770.

Richards, J. A., & Jia, X. (1999). *Remote Sensing Digital Image Analysis. An Introduction*. (3rd ed.). Berlin, Heidelberg, Germany: Springer-Verlag.

Roberts, D. A., Quattrochi, D. A., Hulley, G. C., Hook, S. J., & Green, R. O. (2012). Synergies between VSWIR and TIR data for the urban environment: An evaluation of the potential for the Hyperspectral Infrared Imager (HyspIRI) Decadal Survey mission. *Remote Sensing of Urban Environments.*, *117*, 83–101.

Rojo-Álvarez, J., . Martínez-Ramón, M., Marí Muñoz, J., & Camps-Valls, G. (2017). *Digital Signal Processing with Kernel Methods*. UK: Wiley & Sons. URL: https://www.amazon.com/Digital-Signal-Processing-Kernel-Methods/dp/1118611799.

Roy, D., Wulder, M., Loveland, T., C.E., W., Allen, R., Anderson, M., Helder, D., Irons, J., Johnson, D., Kennedy, R., Scambos, T., Schaaf, C., Schott, J., Sheng, Y., Vermote, E., Belward, A., Bindschadler, R., Cohen, W., Gao, F., Hipple, J., Hostert, P., Huntington, J., Justice, C., Kilic, A., Kovalskyy, V., Lee, Z., Lymburner, L., Masek, J., McCorkel, J., Shuai, Y., Trezza, R., Vogelmann, J., Wynne, R., & Zhu, Z. (2014). Landsat-8: Science and product vision for terrestrial global change research. *Remote Sensing of Environment*, *145*, 154 – 172.

Rudi, A., Calandriello, D., Carratino, L., & Rosasco, L. (2018). On Fast Leverage Score Sampling and Optimal Learning. *arXiv e-prints*, (p. arXiv:1810.13258). arXiv:1810.13258.

Schaum, A., & Stocker, A. (1997). Long-interval chronochrome target detection. In *International Symposium on Spectral Sensing Research* (pp. 1760–1770).

Schölkopf, B., Mika, S., Burges, C. J. C., Knirsch, P., Muller, K.-R., Rätsch, G., & Smola., A. J. (1999). Input space versus feature space in kernel-based methods. *IEEE Transactions on Neural Networks*, *10(5)*, 1000–1017.

Schölkopf, B., & Smola, A. (2002). *Learning with Kernels: Support Vector Machines, Regularization, Optimization and Beyond*. Cambridge, MA, USA: MIT Press Series.

Scott, D. W. (2010). Scott's rule. *WIREs Computational Statistics*, *2(4)*, 497–502.

Shannon, C. E. (2001). A mathematical theory of communication. *ACM SIGMOBILE mobile computing and communications review*, *5(1)*, 3–55.

Shaw, G., & Manolakis, D. (2002). Signal processing for hyperspectral image exploitation. *IEEE Signal processing magazine*, *19(1)*, 12–16.

Shawe-Taylor, J., & Cristianini., N. (2004). *Kernel Methods for Pattern Analysis.*. Cambridge University Press, Cambridge, MA, USA.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis. CRC press* volume 26.

Singh, A. (1989). Digital change detection techniques using remotely-sensed data. *International Journal of Remote Sensing*, *10(6)*, 989–1003.

Snee, R. D. (1983). Regression Diagnostics: Identifying Influential Data and Sources of Collinearity. *Journal of Quality Technology*, *15(3)*, 149–153. doi:10.1080/00224065.1983.11978865.

Stein, D. W. J., Beaven, S. G., Hoff, L. E., Winter, E. M., Schaum, A. P., & Stocker, A. D. (2002). Anomaly detection from hyperspectral imagery. *IEEE Signal Processing Magazine*, *19(1)*, 58–69. doi:10.1109/79.974730.

Sterckx, S., Benhadj, I., Duhoux, G., Livens, S., Dierckx, W., Goor, E., Adriaensen, S., Heyns, W., Van Hoof, K., Strackx, G. et al. (2014). The PROBA-V mission: Image processing and calibration. *International Journal of Remote Sensing*, *35(7)*, 2565–2588.

Stuffler, T., Kaufmann, C., Hofer, S., Förster, K. P., Schreier, G., Mueller, A., Eckardt, A., Bach, H., Penné, B., Benz, U., & Haydn, R. (2007). The EnMAP hyperspectral imager –

An advanced optical payload for future applications in Earth observation programmes. *Acta Astronautica*, *61(1-6)*, 115–120.

Sun, Y., Lei, L., Guan, D., Li, X., & Kuang, G. (2020). Sar image change detection based on nonlocal low-rank model and two-level clustering. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *13*, 293–306.

Theiler, J. (2008). Quantitative comparison of quadratic covariance-based anomalous change detectors. *Applied Optics*, *47(28)*, F12–F26.

Theiler, J. (2013). Spatio-spectral anomalous change detection in hyperspectral imagery. In *2013 IEEE Global Conference on Signal and Information Processing* (pp. 953–956).

Theiler, J. (2014). By definition undefined: Adventures in anomaly (and anomalous change) detection. *6th IEEE Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS)*, .

Theiler, J., & Adler-Golden, S. M. (2008). Detection of ephemeral changes in sequences of images. In *37th IEEE Applied Imagery Pattern Recognition Workshop* (pp. 1–8). doi:10.1109/AIPR.2008.4906469.

Theiler, J., & Grosklos, G. (2016). Cracks in KRX: when more distant points are less anomalous. *8th IEEE Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS)*, (pp. 1–5).

Theiler, J., & Grosklos, G. (2016). Problematic projection to the in-sample subspace for a kernelized anomaly detector. *IEEE Geoscience and Remote Sensing Letters*, *13(4)*, 485–489.

Theiler, J., & Perkins, S. (2006). Proposed framework for anomalous change detection. In *ICML Workshop on Machine Learning Algorithms for Surveillance and Event Detection* (pp. 7–14).

Theiler, J., Scovel, C., Wohlberg, B., & Foy., B. R. (2010). Elliptically Contoured Distributions for Anomalous Change Detection in Hyperspectral Imagery. *IEEE Geoscience and Remote Sensing Letters*, *7(2)*, 271–275.

Theiler, J., & Wohlberg, B. (2012). Local coregistration adjustment for anomalous change detection. *IEEE Transactions on Geoscience and Remote Sensing*, *50(8)*, 3107–3116.

Tibshirani, R. (1996). Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58(1)*, 267–288.

Touati, R., Mignotte, M., & Dahmane, M. (2020). Anomaly feature learning for unsupervised change detection in heterogeneous images: A deep sparse residual model. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *13*, 588–600.

Townshend, J. R. (1994). Global data sets for land applications from the Advanced Very High Resolution Radiometer: an introduction. *International Journal of Remote Sensing*, *15(17)*, 3319–3332.

Ustin, S. (2004). *Remote Sensing for Natural Resource Management and Environmental Monitoring. Manual of Remote Sensing, Volume 4*. New York: John Wiley & Sons.

Volpi, M., Tuia, D., Camps-Valls, G., & Kanevski, M. (2012). Unsupervised Change Detection With Kernels. *IEEE Geoscience and Remote Sensing Letters*, *9(6)*, 1026–1030. doi:10.1109/LGRS.2012.2189092.

Volpi, M., Tuia, D., Kanevski, M., & Camps-Valls, G. (2010). Unsupervised change detection by kernel clustering. In *Image and Signal Processing for Remote Sensing XVI* (p. 78300V). SPIE volume 7830.

Wang, B.-C. (2008). *Digital Signal Processing Techniques and Applications in Radar Image Processing* volume 91. John Wiley & Sons.

Williams, C. K. I., & Seeger, M. (2001). Using the Nyström method to speed up kernel machines. In *Proceedings of the 14th annual conference on neural information processing systems* (pp. 682–688).

Wu, C., Zhang, L., & Du, B. (2017). Kernel Slow Feature Analysis for Scene Change Detection. *IEEE Transactions on Geoscience and Remote Sensing*, *55(4)*, 2367–2384. doi:10.1109/TGRS.2016.2642125.

Yang, T., Li, Y.-f., Mahdavi, M., Jin, R., & Zhou, Z.-H. (2012). Nyström method vs random Fourier features: A theoretical and empirical comparison. In *Neural Information Processing System (NIPS)* (pp. 476–484). volume 25.

Yu, F. X. X., Suresh, A. T., Choromanski, K. M., Holtmann-Rice, D. N., & Kumar, S. (2016). Orthogonal random features. In *Neural Information Processing System (NIPS)* (pp. 1975–1983).

Yuan, Y., Ma, D., & Wang, Q. (2016a). Hyperspectral anomaly detection by graph pixel selection. *IEEE Transactions on Cybernetics*, *46(12)*, 3123–3134.

Yuan, Y., Wang, Q., & Zhu, G. (2016b). Fast hyperspectral anomaly detection via high-order 2-d crossing filter. *IEEE Transactions on Geoscience and Remote Sensing*, *53(2)*, 620–630.

Zhang, C., & Zhang, Z. (2014). Improving multiview face detection with multi-task deep convolutional neural networks. *IEEE Winter Conference on Applications of Computer Vision*, (pp. 1036–1041). doi:10.1109/WACV.2014.6835990.

Zhao, C., Zhao, G., & Jia, X. (2017). Hyperspectral image unmixing based on fast kernel archetypal analysis. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *10*, 331–346.

Zhu, H., Ibrahim, J. G., & Chen, M.-H. (2015). Diagnostic measures for the Cox regression model with missing covariates. *Biometrika*, *102*, 907–923.

Zhu, H., Ibrahim, J. G., & Shi, X. (2009). Diagnostic Measures for Generalized Linear Models with Missing Covariates. *Scandinavian Journal of Statistics*, *36*, 686–712.

Zhu, H., Lee, S.-Y., Wei, B.-C., & Zhou, J. (2001). Case-deletion measures for models with incomplete data. *Biometrika*, *88*, 727–737.