Franklin Open 7 (2024) 100113

Contents lists available at ScienceDirect

Franklin Open

journal homepage: www.elsevier.com/locate/fraope

Shades of green: Unveiling the impact of municipal green bonds on the environment

Marta Campi [a,*], Gareth W. Peters [b], Kylie-Anne Richards [c]

[a] Institut Pasteur, Universite Paris Cite, Inserm, Institut de l'Audition, IHU reConnect, Paris, 75012, France
[b] Department of Statistics and Applied Probability, University of California Santa Barbara, Santa Barbara, 93106-3110, CA, United States of America
[c] Finance Discipline Group, UTS Business School, University of Technology Sydney, PO Box 123 Broadway NSW, Sydney, 2007, Australia

ARTICLE INFO

Keywords:
Green bonds
Kernel principal component analysis
Canonical correlation analysis

ABSTRACT

Green bonds allocate proceeds towards environmentally beneficial projects and sustainable development goals, distinguishing themselves from traditional bonds primarily in the use of proceeds determination. However, investors often find it challenging to assess the carbon reduction potential of these bonds because of the lack of standardised environmental impact reporting. In response to this, our research constructs a unique set of indicators derived from financial and environmental datasets, using multivariate analysis techniques that can accommodate the detection of both linear and non-linear relationships. A novel method combining kernel Principal Component Analysis (kPCA) and Canonical Correlation Analysis (CCA) is applied to detect spatial–temporal cross-correlation in multivariate datasets. This approach handles variable comparability issues and the differential treatment of categorical and numerical variables. A significant finding of this study emerges when this methodology is applied to financial attributes obtained from green bonds issued by municipal agencies (muni bonds), pollution data and environmental (climate) data from nine California counties.

The results of the detailed analysis indicate that there is measurable evidence to indicate relationships between green bond issuance and their use of proceeds for pollution reduction efforts. In particular, the results show a clear and interpretable correlation directly linked to the amount of green bond issuance and the effect this is having on pollution reduction, underscoring the tangible impact these financial instruments have on pollution reduction efforts in California.

Conversely, when it comes to detecting spatial–temporal relationships between the use of proceeds from green bond issuance and positive climate change effects, this is inconclusive from the current studies' analysis. It was found that there were weaker cross-correlation relationships observed between climate and green bond financial data set attributes which is perhaps indicative of the fact that climate change effects take a much longer time frame to occur. As such the findings of the analysis in this regard may not indicate that positive climate change effects are not occurring from green bond initiatives, but rather that the ability to measure detectable improvements to climate with regard to the issuance of green bonds is currently limited and will take longer for such effects to manifest in a statistically detectable fashion from the given data. This is particularly likely to be the case given green bonds are only in their infancy as a financial market, having had the earliest issuance only occurring in the last 15-20 years and only substantial growth in the market over the last 10 years. Therefore, this aspect of the research investigating longer-term climate effects with regard to green bond issuance will take longer to develop.

1. Introduction

The discourse surrounding global warming and climate change has taken centre stage in the decision-making processes of multiple sectors due to their pervasive impacts. The Paris Agreement, established at the 21st Conference of the Parties (COP21) in 2015, underscored the imperative need for a global carbon market and carbon transitions throughout society [1–6]. Despite the multifaceted nature of these challenges, financial markets have emerged as instrumental mechanisms in steering the transition towards low-carbon economies [7–14]. A notable constituent of this financial response is the realm of green finance, wherein green bonds have been particularly well-suited for providing specific financing capacities to companies at a reduced cost,

* Corresponding author.
E-mail address: marta.campi.11@gmail.com (M. Campi).

https://doi.org/10.1016/j.fraope.2024.100113
Received 29 November 2023; Received in revised form 14 May 2024; Accepted 20 May 2024
Available online 23 May 2024

thus ameliorating the economic impact and risk associated with such transitions.

The European Investment Bank (EIB) marked the beginning of this development with the issuance of the first labelled green bond in 2007, called the Climate Awareness Bond [15]. The green bond market has since experienced steady growth, with a plethora of agencies following suit. The unique selling proposition of a green bond, unlike a conventional bond, is its commitment to channel the proceeds from the issuance towards financing green projects, assets, or business activities [16]. Green bonds represent fixed-income financial instruments introduced to raise capital for environmental initiatives through the debt capital market, see a detailed description of green bond instruments, their market dynamics and participants in [17].

The evolution of the green bond market was further shaped by the initiative of a Swedish pension fund consortium in late 2007, which sought to invest in climate change mitigation projects. This led to the World Bank's issuance of the first institutional green bond in November 2008, opening the doors for fixed-income investors to support lending for climate-focused projects. Subsequently, to provide greater structure to this growing market, a consortium of investment banks, including Bank of America Merrill Lynch, Citi, Credit Agricole CIB and JPMorgan Chase & Co., proposed a set of guidelines in January 2014. These guidelines, known as the Green Bond Principles (GBP), provide guidance on the key components of issuing a credible green bond and promote integrity in the green bond market through self-regulation [18]. The GBP was last updated in June 2021, with an additional appendix added in June 2022.

Two challenges arise with current versions of green bond instruments. The first is the often lack of detailed specificity regarding the specific environmental initiatives being targeted or the development goals addressed when it comes to the bond prospectuses outline of the green use of proceeds. A large proportion of green bond prospectus documents refrain from clearly specifying the anticipated environmental impact of the projects to be financed, simply stating that the proceeds will be used for general categories such as wind or solar energy projects. This can culminate in investments that fail to yield the effective environmental impact one would associate with a "green" label. Secondly, there is a scant discussion or description of how such efforts derived from spending the proceeds raised by the bond issuance will be monitored or reported upon with regards to success from the perspective of the environmental impact, pollution reduction or sustainable development goals they seek to provide funding towards. We seek to provide methods to help automatically monitor this sector's success through the development of spatial–temporal multivariate analysis methods.

Nevertheless, green bonds have seen a substantial surge in popularity, solidifying their position as a significant asset class within the global fixed-income market. With the increasing demand for sustainable investment options and the growing urgency to address environmental challenges, the green bond market is poised for further expansion. However, to ensure that green bonds actually deliver their promised environmental benefits, it is crucial to develop robust reporting frameworks and verification standards that allow investors, stakeholders and regulators to make informed decisions and effectively monitor this market with regard to both financial risk and disclosure requirements as well as the environmental performance monitoring and reporting disclosures.

Recent developments within the market, such as the Climate Bonds Initiative's certification for green bonds meeting specific criteria and the European Union's introduction of the EU Green Bond Standard, are commendable strides towards enhanced transparency. However, it is essential to refine these assessment methodologies further to provide comprehensive and accurate assessments of environmental impacts. This includes tracking green bonds throughout their life cycle and addressing potential greenwashing concerns.

The GBP, developed by the International Capital Market Association, and the Climate Bonds Standard, formulated by the Climate Bonds Initiative, have played vital roles in bringing a degree of standardisation to the market by outlining broad project categories contributing to environmental objectives. However, these guidelines are voluntary, leaving issuers to self-label their bonds as green based on guidance from regulators, stock exchanges, and market associations. Regional initiatives, such as the EU's sustainability taxonomy and China's green bond standards, provide further structure to this landscape.

While the global warming and climate change discourse has spurred financial markets towards green finance solutions, several research endeavours have sought to illuminate various facets of this emerging field. One study delves into the spatio-temporal trends of green financing and carbon emissions in the Pearl River Delta, emphasising an increase in green finance alongside a decrease in carbon emissions, with spatial variations indicating higher finance levels in the northeast and higher emissions in the southwest [19]. Additionally, a study explores the dynamic co-movement and risk spillover effects between green bonds and various markets, offering insights for investors and policymakers interested in environmental protection and green investment [20]. Furthermore, a quantitative assessment of ecological vulnerability in the Jianghan Plain sheds light on the determinants of green bond issuance in European Union countries, identifying significant impacts of rating, ESG index, fiscal balance, inflation rate, and population on the volume of green bond issuances [21]. Other research endeavours investigate the coupling relationships between green finance development and industrial transformation [22], analyse the ecological and environmental quality in central China [23], evaluate green bond efficiency in Central and Eastern European countries [24], explore the impact of green bonds on carbon emission intensity in China [25], and assess the efficiency of green bonds in Central and Eastern European countries, highlighting varying yields, durations, and sizes among bonds and examining their comparability [26]. These recent studies collectively contribute to the evolving landscape of green finance, providing valuable insights into its drivers, impacts, and potential for fostering sustainable development.

In this manuscript, we provide a methodological framework that can be adopted to analyse spatial–temporal leading relationships between green bond attributes and pollution reduction or climate change patterns. We believe such a framework and initial analysis is sorely needed in this industry to provide a means to monitor, track and report green bond market pollution or climate mitigation effects at a macro scale. We recognise this is just the starting point for such analysis and acknowledge that many confounding factors may be challenging to resolve, however, we have developed a rigorous first framework that can be built upon over time to refine the monitoring progressively. Whilst this is an early stage of such a monitoring framework, the multivariate methods we developed are far from naive or trivial, they are based upon leading interpretable statistical machine learning methods that are capable to resolve linear and non-linear spatial–temporal relationships between multiple data modalities. This is a very challenging problem to address, and herein lies the statistical novelty of the contributions in this manuscript.

Our research comprehensively examines the evolution, attributes, and impacts of green bonds. We intend to identify the limitations of current methodologies while advancing the development of comprehensive and reliable impact assessment tools. This progression is vital for enhancing the standardisation and transparency of the green bond market. These concerted efforts will be crucial to addressing the growing demand for sustainable investments and promoting improved environmental outcomes.

2. Statistical frameworks methodological context and Green bond case study motivation

A significant gap exists in both the academic literature as well as in industry analysis with regard to monitoring and reporting on the

success of green bond use of proceeds disbursement outcomes with regard to pollution reduction and climate mitigation results. The lack of quantitative environmental impact assessment tools for sovereign, sub-sovereign (including municipal), and state bonds is now a key issue to be addressed as this nascent market grows and eventually matures over the next 10 to 20 years. Therefore, an essential aspect of our research sought to address this absence and contribute to developing methodologies designed explicitly for the quantitative assessment of the environmental impact of these bonds. We intend to substantially improve the evaluation process and effectiveness of the green bond market, promoting greater transparency and environmental benefits.

While significant strides have been made in understanding the multifaceted dynamics of green finance and its intersection with environmental concerns, notable gaps persist, as evidenced by various recent studies (see [19–26]). These works encompass diverse aspects of green finance, ranging from the spatio-temporal evolution of green financing and carbon emissions [19], to the determinants of green bond issuance in European Union countries [21]. While each work sheds light on different facets of green finance, [20,22], and [24] focus on exploring relationships between green finance and various market dynamics, such as stock, crude oil, and gold markets, and the ecological vulnerability of specific regions. Zhao et al. [23], Czech et al. [25] delve into the coupling relationships between green finance and industrial transformation, offering insights into regional development strategies and sustainability initiatives. Moreover, Lee et al. [26] examines the impact of green bonds on carbon emission intensity in China, providing valuable insights into the effectiveness of green investment mechanisms. Although the focus of these studies varies, they collectively underscore the growing importance of green finance in addressing environmental challenges and transitioning towards sustainable economies. In contrast, our research aims to develop a novel methodological framework specifically tailored for assessing the environmental impact of green bond use of proceeds, filling a crucial gap in quantitative assessment tools for evaluating the success of green bond initiatives in pollution reduction and climate mitigation.

Our decision to focus this research on the U.S. municipal green bond market is informed by two primary factors: the substantial carbon footprint of the U.S. and the size, financial disclosure reporting transparency and potential of its municipal green bond market. As one of the leading global contributors to greenhouse gas emissions, the U.S. is crucial in mitigating climate change. Deepening our understanding of the U.S. municipal green bond market could illuminate strategies to lower carbon emissions and expedite the shift towards a low-carbon economy.

In terms of market size, the U.S. municipal green bond market comprises a significant fraction of the global green bond market, with a total value exceeding hundreds of billion dollars. This market provides crucial funding for numerous public projects with significant environmental implications. States and municipalities have used Green bonds extensively to finance projects to improve sustainability and counter climate change. For example, California, the largest state issuer of municipal bonds, has used green bonds to fund various projects, from renewable energy generation to improvements in water infrastructure and sustainable transportation systems. Other states, including New York and Massachusetts, have similarly leveraged green bonds for climate change mitigation initiatives.

Researching this market can provide valuable insight into how green bonds can effectively promote environmental sustainability. We can better understand the market's function and impact by examining specific cases from different states. This localised approach is critical to developing tailored strategies and policies considering various U.S. regions' unique circumstances and needs.

The scope of our study narrows to focus further on the state of California for several reasons. Firstly, California has issued many green bonds, resulting in a wealth of data for analysis. These data sets are also publicly available, facilitating access and replication of results. Furthermore, evidence indicates that pollution and climate change impacts are particularly severe in California. The state leads the nation in levels of ozone pollution, with several of its cities ranking among the most polluted in the American Lung Association's "State of the Air" report [27]. California also faces severe climate change-related challenges, such as frequent wildfires, persistent droughts, and rising sea levels [28]. This context underscores the pressing need for environmental risk monitoring in the state. Furthermore, the number of monitors for climate and pollution variables is much higher in California than in the rest of the United States. We provide further information about this in the Supplementary Information.

This work aims to reveal, identify and measure the environmental impact of green project disbursements in areas that are highly polluted and highly populated through the use of three datasets: green bonds, pollution, and climate data. The achievement of this research required two main contributions.

First, we collected relevant variables for the three types of data information and engineered appropriate statistical features from available reported spatial–temporal attributes. This task required advanced data processing, data cleaning, and data wrangling, leading to the construction of three data sets available at https://github.com/mcampi111 that can be reused for further research purposes. The three data sets have been compiled as follows, based on reputable leading data providers:

1. Pollution data has been constructed by downloading information from the US Environmental Protection Agency website[1];
2. Climate data set has been constructed by extracting variables from one of the National Oceanic and Atmospheric Administration (NOAA) data sets, specifically the Global Surface Summary of the Day (GSOD) data set[2];
3. Green bonds dataset has been collected through the Bloomberg Terminal and provides information on municipal green bonds issued within the US State of California.

The second element of our work involved developing a methodology to study potential relationships between green bond issuance and these financial instruments' multifaceted attributes and the change in pollution and climate over space and time in areas where the green bond use of proceeds is being deployed for potential sustainable development goals. In doing so, we observe that numerous aspects must be considered. First, the procedure has to deal with the non-stationary and non-linear nature of the relationships that may be present in data over space and time. Second, one must treat a multi-modal data source. This involves determining appropriate attributes to record in space and time and to integrate these into non-linear factor extractions that can be studied for associations in space and time and will be informative on statistical associations and dependence relationships between each data modality.

Thirdly, the results should be interpretable in terms of each modality's measure attributes. As this will ensure that one can attempt to measure the environmental impact of green bond issuance. In this regard, we first look for robust data feature extraction methods applied over individual data sets (modalities). Then subsequently, we study statistical associations between each modality based on the leading extracted features from each modality. This allows us to quantify the statistical association between the computed modes of variation across the data sets. In this way, accurate spatio-temporal correlations of green bonds and pollution/climate variations can be robustly detected. Fourth, the correlation might depend on time or spatial features, and the implemented method must be able to capture such information. Fifth, the statistical interpretation of the method must be provided so

[1] https://www.epa.gov/

[2] https://www.ncei.noaa.gov/access/metadata/landing-page/bin/iso?id=gov.noaa.ncdc:C00516

that practitioners can benefit and conduct more efficient data decision-making processes. Such tasks are attained by combining two methods known as Kernel Principal Component Analysis (kPCA, see [29]) and Canonical Correlation Analysis (CCA, [30–32]).

kPCA is a widely used machine learning technique corresponding to the non-linear version of standard Principal Component Analysis (PCA, see [33]) that converts a certain number of potentially correlated variables into a set of uncorrelated Principal Components (PCs), capturing the variability of the underlying data set. If the PCs are non-linearly related to the input variables, this technique fails and provides a misleading explanation of the data variability. When this is the case, kPCA can be used instead. kPCA belongs to the class of kernel methods [29,34] whose idea is to map the existing data set into a new space, called the feature space which can be a function space, where linear algorithms are applied again. The advantage of this approach is that one does not have to know explicitly the feature map functions explicitly as they can be obtained implicitly via the specification of a kernel and application of the so-called "kernel trick". The advantage of this technique is the robustness of the kernel PCs (kPCs), identifying the variability of the original data by handling its non-linearity and non-stationarity.

In this work, the kPC functions from each data modality are the input of the CCA multivariate spatial–temporal dependence analysis. CCA is a statistical method that models the association among two multivariate data sets by providing a set of canonical variates corresponding to orthogonal linear combinations of the variables within each data set that exhibit maximum correlation. Hence, CCA identifies new variables that maximise the interrelationships between two data set modalities, in contrast to the new variables of PCA describing the internal variability within one data set modality. The application of the CCA method in the manner we propose is designed specifically to capture the intra-data set modalities' statistical relationships. The novelty of this work from a statistical perspective will be to use the CCA to identify cross-correlation over the robust kPCs extracted by the different data set modalities.

We note that our proposed method of first applying kPCA to each data modality and then CCA between each resultant data modalities kPC functions is distinct from other works who seek to develop kernel CCA methods, see [35–43]. Unlike these methods, in our approach we maintain greater interpretability of all discovered spatial–temporal relationships between data modalities through our proposed two-stage approach of inter-modality kPCA followed by intra-modality CCA. In application of this method we can consider several multivariate input data sets carrying spatial information, for example, data sets collecting pollution variables in each Californian county over time, data sets for climate variables in each county of California over time and data sets of green bond financial variables of the same California counties over time. Hence, one will have one multivariate triple of data modalities per county over time. We will then transform each modality into a set of non-linear spatial–temporal explanatory factors via kPCA per modality. The choice of kernels we consider allow us to treat multiple types of data including time series, ordinal and categorical data that may arise in each data modality. Having extracted the factor representations for each data modality, as kPC's for instance the first kernel principal component extracted by the pollution data of all the counties and the first kernel principal component extracted by the financial data set of all the counties, we then apply CCA methods to study their spatial–temporal relationships between each data sets modality.

There are several advantages to our proposed methodology. The first is that data attributes from each coordinate dimension can be irregularly sampled in time and space and still be incorporated into this analysis since the kPCA will provide an output function approximation of the kPC's that can be then evaluated on a common mesh of space and time points. In this way, the CCA applied across modalities of data sets when applied to the kPC's will be on a common space–time grid.

We claim that if there is an association between green bond variables and pollution variables (for example), we can interpret such an association as evidence that the green initiatives funded by green bond proceeds are associated with climate mitigation impacts. The constructed methodology deals with variables whose processes are observed over time and will average this information by providing an instantaneous spatial correlation. As a result, we will be able to observe the time relationship between the green bond variables and the pollution variables, as a delay in the impact of the green project is expected. The methodology will first extract the kernel principal components (kPCs) for every data set to identify leading factors that efficiently select the spectral content of the original data in an automated fashion based on the size of the eigenvalues. The kPCs will then be fed to the CCA to observe cross-correlation between the leading variations of the original data to robustly define the environmental impact of green projects in the different counties of California. Arguments for combining these two techniques are as follows. If one applies the CCA directly to the raw data, the captured information is nothing more than the cross-correlation between the data sets. The critical issue with such a standard approach is the complex structural information of the underlying data, which is non-linear and non-stationary. Hence, the CCA would detect a great deal of noise and erratic association, which pollutes final decision-making processes. The role of kPCA in this instance is to effectively detect the variability of the underlying data and discard irrelevant information. Furthermore, such a technique will provide the relevant spectral components of the data in an automated fashion according to the most dominant eigenvalues, i.e. eigenmodes. This practise could be done by applying, for example, spectral truncation, which is, however, highly difficult since identifying which spectral components to retain for an efficient final representation is challenging in practise. The role of the CCA is now finalised to the task of interest, i.e. it will then focus on the cross-correlation between the dominant marginal eigenmodes.

2.1. Contributions, notation and organisation of the paper

The contributions of this work can be split into notional, methodological and data applied and are given as follows:

- The first contribution is given by a quantitative definition of the environmental impact of a green bond. Such a concept is, in general, highly debated within the green finance community and, therefore, particularly needed. In this respect, the second relevant point is to provide reliable statistical indicators capturing variability with leading factors to measure such impact, which practitioners strongly require of the financial markets in general. The kPCs will act as such, and we will show that diversified information will be detected depending on the dataset and the KPC number.
- The second contribution of this work is the combination of the kPCA and the CCA to detect cross-correlation amongst datasets in non-stationary settings, which is a required tool for any application dealing with such an issue. We show that this approach strongly empowers the desired findings with respect to traditional linear PCA combined with CCA or with CCA applied to the data only. Hence, this will be a robust version of CCA over non-linear and non-stationary datasets.
- The considered datasets contain multiple types of input variables, i.e. numerical, categorical, etc. Therefore, the treatment of different sources of data is an issue that often affects kernel methods in general. We overcome such an issue by employing the Jaccard Distance and embedding the contribution of the categorical data through the Jaccard kernel allowing for multi-modality kPCA, which is highly needed in the analysis of environmental and financial data.

- Another relevant contribution is the engineering of ad hoc variables and features required to identify the environmental impact over time and space accurately. This is necessary since the considered data differs regarding observational timestamps and spatial recording monitors. Hence, feature engineering is required to identify relevant information carrying the variability, which is informative to reveal the behaviour of the underlying data.

We will denote the three constructed datasets, each of different modality, by $\mathbf{X}^1_{N_1\times D_1}$, $\mathbf{X}^2_{N_2\times D_2}$ and $\mathbf{X}^3_{N_3\times D_3}$, where the first one represents the pollution dataset, the second one the climate dataset and the third one the financial green bond dataset respectively. The first two datasets collect attributes related to pollution or climate (accurately described in Section 5) from different pollution or climate stations which are selected according to their distance from the counties of interest in California. We will adopt the notation for each data set where $\mathbf{X}^i_{N_i\times D_i}$ for all $i \in \{1,2,3\}$, for the first two data sets the index $N_i = T_i \times S$, where T_i corresponds to the number of time samples collected at s locations and S the total number of locations, i.e. the counties studied in California and for the third data set N_i will correspond to the number of green bonds issued by each county over the time period of interest. We denote by D_i the attribute signals that have been observed for the ith data modality. For instance if $i = 1$ then the attributes are pollution related variables being measured such as carbon dioxide, air quality (see details in Section 5); if $i = 2$ then the attributes are observed climate signals such as total precipitation, temperature, (see details in Section 5); and if $i = 3$ then the attributes are observed financial green bond attributes such as coupon rates, maturities, industry of the green bond, use of proceeds green initiatives label (see details in Section 5).

Feature extraction is performed using linear and non-linear projection methods. The set of such projection bases i.e. PCs and kPCs will be extracted by splitting the dataset according to each considered county in California, hence we will have for the ith data modality $\mathbf{X}^i_{N_i\times D_i} = [\mathbf{x}^{i,1}_{T_i\times D_i}, \mathbf{x}^{i,2}_{T_i\times D_i}, \dots, \mathbf{x}^{i,S}_{T_i\times D_i}]$, where subscript remain as explained previously and the upper subscripts denote the data modality index i followed by the county index $s \in \{1,\dots,S\}$.

The paper is organised as follows: first, a section reviewing the statistical methods of PCA, kPCA and the CCA is provided. This includes a review of the out-of-sample and pre-image problems for the kPCA method and the introduction of the required notations for each method. Second, a methodology section is presented, with the introduction of the novel method proposed in this manuscript termed kPCA-CCA used for analysis of several multi-attribute (i.e. multivariate) spatial–temporal data sets to analysis associations between attributes from each each data modality in space and time. This is followed by a detailed comparison between the reference method for linear methods based on PCA-CCA methodology and the extended method to treat non-linearity and non-stationarity achieved by the kPCA-CCA framework. Subsequently, the results for the data and the experiments, are presented. Finally, the paper discussion along with the conclusions are provided. Supplementary Information accompany the manuscript which provide detailed explanations of a practical nature regrading data processing and kernel preparation. All code and data is provided for reproducibility in the github repository with paper name under https://github.com/mcampi111.

3. Spatial-temporal methods to assess Green bond disbursements: PCA, kPCA and CCA

This section briefly reviews the concept of PCA and the kPCA by showing the steps required to extract the PCs and the kPCs, respectively. For the kPCA, we distinguish between the cases of knowing the feature mapping ϕ and not knowing it. Then, the out-of-sample and the pre-image problems are reviewed. The last part of the section is dedicated to the presentation of the CCA, its proposed model and the constrained maximisation problem that must be solved. Remark that, while the PCA and the kPCA look for the internal variability of a given set of variables in a linear and non-linear fashion, the CCA is a procedure searching for cross-correlation or interrelationships between two sets of data.

The methodologies introduced in Section 4 extract PCs and kPCs on the three datasets split by counties, and this will be presented and explained within Section 4. For simplicity and without any loss of generality, we review the following methods by referring to a general input matrix $\mathbf{X}_{N\times D}$.

3.1. PCA for intra data modality to capture linear variability

We present the PCA framework, which will be extended to its non-linear version in the following subsection. Consider N samples $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N \in \mathcal{X} \subset \mathbb{R}^{1\times D}$ the set of data observed in the input space. When stacked by row, the data matrix is denoted $\mathbf{X}_{N\times D}$ with each observation a row $\mathbf{x}_n = [x_{n,1}, \dots, x_{n,D}]_{1\times D}$. We recall in this section the concept of PCA which searches for basis vectors of unit length and orthogonal to each other, hence orthonormal bases, to re-express a given dataset $\mathbf{X}$ under projection as follows:

$$\mathbf{X}_{N\times D}\mathbf{W}_{D\times D} = \mathbf{L}_{N\times D} \tag{1}$$

where $\mathbf{W}$ is a $D \times D$ matrix and denotes a linear projection. The columns of $\mathbf{W}$ are the new basis vectors which, by construction, provide $\mathbf{W}^\intercal\mathbf{W} = \mathbb{I}_D$, and express rows of $\mathbf{L}$. Such a re-expression makes the PCA a method for lowering the redundancy in the $\mathbf{X}$ data set. It can be formally written for i, j columns of $\mathbf{L}$ as

$$[\mathbf{L}]^\intercal_{\cdot,i}[\mathbf{L}]_{\cdot,i} = [\mathbf{W}]^\intercal_{\cdot,i}\mathbf{S}_X[\mathbf{W}]_{\cdot,i} \quad \text{and} \quad [\mathbf{L}]^\intercal_{\cdot,i}[\mathbf{L}]_{\cdot,j} = [\mathbf{W}]^\intercal_{\cdot,i}\mathbf{S}_X[\mathbf{W}]_{\cdot,j} = 0$$

where $\mathbf{S}_X = \mathbf{X}^\intercal\mathbf{X}$ represents the sample estimate of the population covariance. PCA extracts its basis functions through a procedure which directly acts on $\mathbf{S}_X$. In practice, it seeks a linear combination of Eq. (1) that maximises the overall variance of $\mathbf{L}$ given by $\mathbf{S}_L = \mathbf{L}^\intercal\mathbf{L}$. The solution to the problem is found by a maximiser with the following Lagrangian expression given by

$$Q(\mathbf{W}) = \mathbf{W}^\intercal\mathbf{S}_X\mathbf{W} - \boldsymbol{\Lambda}\left(\mathbf{W}^\intercal\mathbf{W} - \mathbb{I}_D\right)$$

for $\boldsymbol{\Lambda}_{D\times D}$ being a diagonal $D \times D$ matrix with Lagrangian coefficients. One can then solve the optimisation problem by differentiating and finding the turning points to solve for the roots of the quadratic form of this objective function as follows

$$\frac{\partial Q}{\partial \mathbf{W}} = 2\mathbf{S}_X\mathbf{W} - 2\boldsymbol{\Lambda}\mathbf{W} = 0 \implies \mathbf{S}_X\mathbf{W} = \boldsymbol{\Lambda}\mathbf{W}$$

It is possible to observe that $\mathbf{W}$ is a matrix in which columns are eigenvectors of $\mathbf{S}_X$, whereas $\boldsymbol{\Lambda}$ is a matrix of corresponding eigenvalues with the number of the non-zero elements equal to the rank of $\mathbf{S}_X$. The columns of $\mathbf{L}$ are indeed orthogonal since

$$[\mathbf{L}]^\intercal_{\cdot,i}[\mathbf{L}]_{\cdot,j} = [\mathbf{W}]^\intercal_{\cdot,i}\mathbf{S}_X[\mathbf{W}]_{\cdot,j} = [\mathbf{W}]^\intercal_{\cdot,i}\lambda_j[\mathbf{W}]_{\cdot,j} = \lambda_j[\mathbf{W}]^\intercal_{\cdot,i}[\mathbf{W}]_{\cdot,j} = 0$$

It can be easily proven that $\mathbf{L}$, defined by $\mathbf{W}$, the eigenvectors of $\mathbf{S}_X$, maximises the total trace of $\mathbf{S}_L$, its determinant and maximises the Euclidean distance between the columns of $\mathbf{L}$. Furthermore, the representation minimises the mean square error between the observations and its projection.

3.2. Kernel kPCA for non-linear variability of intra data modality

This subsection presents the non-linear kernel version of PCA known as kPCA, showing the steps required for its derivation, in both cases when the feature mapping is known and unknown. Afterwards, the out-of-sample problem is discussed. This corresponds to the case in which a new sample data point $\mathbf{x}^\star$ is considered and must be mapped onto the kernel space identified through the kPCA. Such a procedure is highly relevant for the development of our methodology proposed in the following section and lends importantly to the interpretation of

results obtained. The last part presented in this subsection is the pre-image problem, meaning the exercise of projecting the obtained feature points back into the original input space. We will rely on this element of kPCA for the hyperparameters grid search performed to identify the optimal kPCs in our experiments.

3.2.1. Background and main objectives of the kPCA

kPCA is utilised when one seeks to detect non-linear and non-stationary features characterising each of the data sets considered. Assume in the context of kernel kPCA that $\boldsymbol{\phi} : \mathcal{X} \to \mathcal{F}$, represents a mapping from the observed input space to the feature space $\mathcal{F} \subset \mathbb{R}^P$ that is typically non-linear, such that

$$\boldsymbol{\Phi}_{N\times P} = \begin{bmatrix} \boldsymbol{\phi}(\mathbf{x}_1) \\ \vdots \\ \boldsymbol{\phi}(\mathbf{x}_N) \end{bmatrix}_{N\times P} = \begin{bmatrix} \phi_1(\mathbf{x}_1) & \dots & \phi_P(\mathbf{x}_1) \\ \vdots & \ddots & \vdots \\ \phi_1(\mathbf{x}_N) & \dots & \phi_P(\mathbf{x}_N) \end{bmatrix}_{N\times P}.$$

Denote $\mathbf{C} = \boldsymbol{\Phi}^\intercal\boldsymbol{\Phi}$ the $P \times P$ positive definite matrix representing the sample estimate of the covariance matrix of $\boldsymbol{\Phi}_{N\times P}$ in the feature space. One cell of the matrix $\mathbf{C}$ at the ith row and the jth column is given as

$$C_{i,j} = [\phi_i(\mathbf{x}_1), \dots, \phi_i(\mathbf{x}_N)] \begin{bmatrix} \phi_j(\mathbf{x}_1) \\ \vdots \\ \phi_j(\mathbf{x}_N) \end{bmatrix} = \sum_{n=1}^{N} \phi_i(\mathbf{x}_n)\phi_j(\mathbf{x}_n)$$

and represents the covariance between the ith feature function ϕ_i and jth feature function ϕ_j within the feature space $\mathcal{F}$ for $i, j = 1, \dots, P$ across all available samples of data.

Furthermore, we also consider the Gram Matrix which is characterised by a Mercer kernel covariance operator denoted $k : \mathcal{X}\times\mathcal{X} \to \mathbb{R}$, known as the kernel function that defines the inner product in the feature space $\mathcal{F}$ that is given as

$$\begin{aligned} k(\mathbf{x}_n, \mathbf{x}_m) = \boldsymbol{\phi}(\mathbf{x}_n)\boldsymbol{\phi}(\mathbf{x}_m)^\intercal &= [\phi_1(\mathbf{x}_n), \dots, \phi_P(\mathbf{x}_n)] \begin{bmatrix} \phi_1(\mathbf{x}_m) \\ \vdots \\ \phi_P(\mathbf{x}_m) \end{bmatrix} \\ &= \sum_{p=1}^{P} \phi_p(\mathbf{x}_n)\phi_p(\mathbf{x}_m) \end{aligned} \tag{2}$$

for $n, m \in \{1, \dots, N\}$. Then we define $\mathbf{K}$ the $N \times N$ Gram Matrix such that $\mathbf{K}_{N\times N} = \boldsymbol{\Phi}\boldsymbol{\Phi}^\intercal$.

The kPCA projects $\boldsymbol{\Phi}_{N\times P}$ onto uncorrelated components, possibly of smaller dimensionality. This is achieved by expressing each point $\boldsymbol{\phi}_n \in \mathcal{F}$ as a linear combination of $Q \le P$ orthogonal basis vectors $\{\mathbf{v}_1, \dots, \mathbf{v}_Q\}$ of dimension P, where $\mathbf{v}_q \in \mathbb{R}^P$, and uncorrelated weight coefficient scores $\{\boldsymbol{a}_1, \dots, \boldsymbol{a}_Q\}$, where $\boldsymbol{a}_q \in \mathbb{R}^N$, such that the following representation holds

$$\boldsymbol{\phi}_n = \boldsymbol{\phi}(\mathbf{x}_n) = \sum_{q=1}^{Q} a_{n,q}\mathbf{v}_q$$

If one considers the matrix $\mathbf{A}_{N\times Q}$ and $\mathbf{V}_{Q\times P}$ such that

$$\mathbf{A}_{N\times Q} = [\boldsymbol{a}_1, \dots, \boldsymbol{a}_Q] = \begin{bmatrix} a_{1,1} & \dots & a_{1,Q} \\ \vdots & \ddots & \vdots \\ a_{N,1} & \dots & a_{N,Q} \end{bmatrix}_{N\times Q} \quad \text{and}$$

$$\mathbf{V}_{Q,P} = \begin{bmatrix} \mathbf{v}_1 \\ \vdots \\ \mathbf{v}_Q \end{bmatrix} = \begin{bmatrix} v_{1,1} & \dots & v_{1,P} \\ \vdots & \ddots & \vdots \\ v_{Q,1} & \dots & v_{Q,P} \end{bmatrix}_{Q\times P}$$

given the assumptions of uncorrelated columns of $\mathbf{A}_{N\times Q}$ and orthonormality of rows in $\mathbf{V}_{Q\times P}$, we obtain

$$\mathbf{A}^\intercal\mathbf{A} = \boldsymbol{\Lambda}_{Q\times Q} \quad \text{and} \quad \mathbf{V}\mathbf{V}^\intercal = \mathbb{I}_Q$$

The representation that we are trying to find is then defined in matrix form as

$$\boldsymbol{\Phi}_{N\times P} = \mathbf{A}_{N\times Q}\mathbf{V}_{Q\times P}$$

3.2.2. kPCA when the feature mapping ϕ is known

If the mapping ϕ is known, the covariance matrix can be defined and, therefore, standard eigenvalue decomposition on $\mathbf{C}$ can be applied to obtain $\mathbf{V}$ since

$$\mathbf{C} = \boldsymbol{\Phi}^\intercal\boldsymbol{\Phi} = \mathbf{V}^\intercal\mathbf{A}^\intercal\mathbf{A}\mathbf{V} = \mathbf{V}^\intercal\boldsymbol{\Lambda}\mathbf{V}$$

and by multiplying $\mathbf{V}$ on the left of both sides, one obtains the eigen-equation given as

$$\mathbf{V}\mathbf{C} = \boldsymbol{\Lambda}\mathbf{V}$$

The rows of $\mathbf{V}$ are Q eigenvectors of $\mathbf{C}$ and $\boldsymbol{\Lambda}_{Q\times Q}$ is the matrix with corresponding eigenvalues. Then, since $\mathbf{V}\mathbf{V}^\intercal = \mathbb{I}_Q$, we have

$$\boldsymbol{\Phi}\mathbf{V}^\intercal = \mathbf{A}\mathbf{V}\mathbf{V}^\intercal \implies \boldsymbol{\Phi}\mathbf{V}^\intercal = \mathbf{A} \tag{3}$$

and $\mathbf{A}$ is the new representation of $\boldsymbol{\Phi}$ that could be of lower dimension and represents the matrix of the Principal Components computed in the feature space.

3.2.3. Unknown feature mapping ϕ

The feature mapping ϕ is usually unknown and must be learnt or specified in the kernel modelling exercise. Therefore, the covariance matrix $\mathbf{C}$ cannot be computed or might require a high computational cost given that $\mathbf{C}$ is, in general,highly dimensional. As a result, the projection $\mathbf{V}$ is not explicitly known. One way to solve this problem is given by utilisation of a kernel function $k(\cdot,\cdot)$, as given in Eq. (2). Next, we show that the principal components $\mathbf{A}$ in the feature space, i.e. the kPCs, and their corresponding variances stored on the diagonal of $\boldsymbol{\Lambda}$ can be obtained by using only the Gram Matrix $\mathbf{K}_{N\times N}$ defined through the inner product of the kernel. We can substitute $\mathbf{C}$ with $\boldsymbol{\Phi}^\intercal\boldsymbol{\Phi}$ and multiply on the left by $\mathbf{V}$, we obtain

$$\begin{aligned} \mathbf{C} &= \mathbf{V}^\intercal\boldsymbol{\Lambda}\mathbf{V} \\ \mathbf{V}\boldsymbol{\Phi}^\intercal\boldsymbol{\Phi} &= \underbrace{\mathbf{V}\mathbf{V}^\intercal}_{=\mathbb{I}_Q}\boldsymbol{\Lambda}\mathbf{V} \end{aligned}$$

Then, by multiplying both sides times $\boldsymbol{\Phi}^\intercal$ and considering $\mathbf{K}_{N\times N} = \boldsymbol{\Phi}\boldsymbol{\Phi}^\intercal$, one can then derive

$$\begin{aligned} \mathbf{V}\boldsymbol{\Phi}^\intercal\underbrace{\boldsymbol{\Phi}\boldsymbol{\Phi}^\intercal}_{=\mathbf{K}_{N\times N}} &= \boldsymbol{\Lambda}\mathbf{V}\boldsymbol{\Phi}^\intercal \\ \mathbf{A}^\intercal\mathbf{K} &= \boldsymbol{\Lambda}\mathbf{A}^\intercal \end{aligned}$$

as $\boldsymbol{\Phi}\mathbf{V}^\intercal = \mathbf{A}$. The matrix $\mathbf{A}_{N\times Q}$ are almost eigenvectors of the gram matrix $\mathbf{K}$ as they are not orthonormal yet (only orthogonal so far) since

$$\mathbf{A}^\intercal\mathbf{A} = \mathbf{V}\boldsymbol{\Phi}^\intercal\boldsymbol{\Phi}\mathbf{V}^\intercal = \mathbf{V}\mathbf{C}\mathbf{V} = \boldsymbol{\Lambda}$$

By rescaling both sides of the equation by the square root of $\boldsymbol{\Lambda}$ we obtain

$$\begin{aligned} \boldsymbol{\Lambda}^{-\frac{1}{2}}\mathbf{A}^\intercal\mathbf{K} &= \boldsymbol{\Lambda}^{-\frac{1}{2}}\boldsymbol{\Lambda}\mathbf{A}^\intercal \\ \mathbf{Z}^\intercal\mathbf{K} &= \boldsymbol{\Lambda}\mathbf{Z}^\intercal \end{aligned}$$

We have obtained orthonormal eigenvectors $\mathbf{Z}_{N\times Q} = \mathbf{A}\boldsymbol{\Lambda}^{-\frac{1}{2}} = \boldsymbol{\Phi}\mathbf{V}^\intercal\boldsymbol{\Lambda}^{-\frac{1}{2}}$ since $\mathbf{Z}^\intercal_{Q\times N}\mathbf{Z}_{N\times Q} = \boldsymbol{\Lambda}^{-\frac{1}{2}}\boldsymbol{\Lambda}\boldsymbol{\Lambda}^{-\frac{1}{2}} = \mathbb{I}_Q$. Therefore, by taking the eigen-decomposition of the gram matrix $\mathbf{K}$ we obtain the matrices $\mathbf{Z}_{N\times Q}$ and $\boldsymbol{\Lambda}$ and we will need to rescale them to obtain the matrix $\mathbf{A}_{N\times Q}$, which is the matrix with the principal components (corresponding to the representation of sample points from $\mathcal{F}$ in the new feature space projected by $\mathbf{V}_{Q\times P}$). These are usually referred to as the kernel Principal Components, kPCs, since extracted through the use of the kernel function only and defined within the feature space.

3.2.4. The out-of-sample problem with unknown ϕ

This section aims to show how to evaluate the unknown feature map ϕ at any point $\mathbf{x}^\star$ which is not in the training set, i.e. $\mathbf{x}^\star \notin \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N\}$. This is achieved by defining a new sample $\phi(\mathbf{x}^\star)$ through the decomposition of the gram matrix $\mathbf{K}$, which is based on the samples $\{\phi(\mathbf{x}_1), \dots, \phi(\mathbf{x}_N)\} \in \mathcal{F}$, and is evaluated through the learnt kernel map k. Recall that $\boldsymbol{\Phi} = \mathbf{AV}$, so multiplying both sides by $\mathbf{A}^\intercal$ and rearranging produces

$$\mathbf{V}_{Q\times P} = \boldsymbol{\Lambda}^{-1}\mathbf{A}^\intercal\boldsymbol{\Phi}.$$

Then by setting $\mathbf{W}_{N\times Q} = \mathbf{A}_{N\times Q}\boldsymbol{\Lambda}^{-1}_{Q\times Q}$, we achieve the formulation of the eigenvectors $\mathbf{V}$ as an expression of the feature samples $\boldsymbol{\Phi}$ given as $\mathbf{V}_{Q\times P} = \mathbf{W}^\intercal_{Q\times N}\boldsymbol{\Phi}_{N\times P}$, which, for an individual vector $\mathbf{v}_q$, corresponds to

$$\mathbf{v}_q = \sum_{n=1}^{N} w_{n,q}\boldsymbol{\phi}(\mathbf{x}_n) = \sum_{n=1}^{N} \frac{\langle \mathbf{v}_q, \boldsymbol{\phi}(\mathbf{x}_n)\rangle}{\langle \boldsymbol{\phi}(\mathbf{x}_n), \boldsymbol{\phi}(\mathbf{x}_n)\rangle}\boldsymbol{\phi}(\mathbf{x}_n)$$

On the other hand, given $\boldsymbol{\Phi} = \mathbf{AV}$, we can write an individual vector $\boldsymbol{\phi}(\mathbf{x}_n)$ as

$$\boldsymbol{\phi}(\mathbf{x}_n) = \sum_{n=1}^{N} a_{n,q}\mathbf{v}_q = \sum_{n=1}^{N}\langle \mathbf{v}_q, \boldsymbol{\phi}(\mathbf{x}_n)\rangle\mathbf{v}_q$$

By using these last two considerations, we observe that the projection of $\boldsymbol{\phi}(\mathbf{x}_m)$ is

$$a_{m,q} = \boldsymbol{\phi}(\mathbf{x}_m)\mathbf{v}_q^\intercal = \boldsymbol{\phi}(\mathbf{x}_m)\sum_{n=1}^{N} w_{n,q}\boldsymbol{\phi}(\mathbf{x}_n)^\intercal = \sum_{n=1}^{N} w_{n,q}\boldsymbol{\phi}(\mathbf{x}_m)\boldsymbol{\phi}(\mathbf{x}_n)^\intercal$$
$$= \sum_{n=1}^{N} w_{n,q}k(\mathbf{x}_m, \mathbf{x}_n)$$

Given the above, by multiplying times $\mathbf{v}_q$ both sides, this will simplify to an expression that will hold for any $\boldsymbol{\phi}(\mathbf{x}^\star) \in \mathcal{F}$ allowing for its definition only through the kernel function and the eigendecomposition of $\mathbf{K}$, that is

$$\phi(\mathbf{x}^\star) = \sum_{n=1}^{N} w_{n,q}k(\mathbf{x}^\star, \mathbf{x}_n)\mathbf{v}_q \tag{4}$$

3.2.5. The pre-image problem

In this section, we provide a brief review of the pre-image problem and the solution that we adopted in this work. Once the sample points are mapped into the feature space $\mathcal{F}$, then it is often the case that one wants to map them back to the input space $\mathcal{X}$. This exercise is identified in the literature as the pre-image problem and affects kernel methods in general, and, in this case, our interest in the Kpca version. We review this concept since we will be using such a method for the hyperparameter learning procedure.

The pre-image problem consists in finding the counterpart of $\phi(\mathbf{x})$ back in the input space $\mathcal{X}$, i.e a point $\tilde{\mathbf{x}}$ such that $\phi(\tilde{\mathbf{x}}) = \phi(\mathbf{x})$. However, the map ϕ is usually non-linear and, therefore, might not be invertible uniquely. This is indeed an ill-posed problem where one seeks an approximate solution denoted as $\hat{\mathbf{x}} \in \mathcal{X}$ whose map $\phi(\hat{\mathbf{x}})$ is as close as possible to $\phi(\mathbf{x})$.

The pre-image problem then can be reformulated and interpreted as finding the approximation $\hat{\mathbf{x}}$ through solving the following optimisation problem

$$\hat{\mathbf{x}} = \text{argmin}_{\mathbf{x}\in\mathcal{X}}\|\phi(\mathbf{x}) - \phi(\hat{\mathbf{x}})\|^2_{\mathcal{F}}.$$

Several solutions have been proposed in the literature, see [44] and summary in the Supplementary Information. In this work we adopt the solution from Bakır et al. [45]. Since the function $\phi(\cdot)$ is defined on a vector space it can be represented vector-wise through any orthonormal basis spanning the subspace where it lies (the feature space $\mathcal{F}$). The orthonormal basis considered in this work is the kPCA. We have introduced the notation for the projection of any input $\mathbf{x}$ in previous section and recall it here as $a_q = \phi(\mathbf{x})\mathbf{v}_q^\intercal$. If we consider the projection on every q-th axes we obtain

$$P_k\phi(\mathbf{x}) = \phi(\mathbf{x})\mathbf{V}^\intercal = [\phi(\mathbf{x})\mathbf{v}_1^\intercal, \dots, \phi(\mathbf{x})\mathbf{v}_Q^\intercal]_{1\times Q}$$

where the operator P_k highlights that such a projection is induced through the kernel $k(\cdot,\cdot)$. If the considered kernel is universal and characteristic, as the sample size increases, and therefore the number of obtained kPCA eigenvectors increases, then one would expect that this pointwise projection representation of ϕ will become more accurate, i.e. that $\phi(\mathbf{x}) \approx P_k\phi(\mathbf{x})$. Hence, we look for an approximation $\hat{\mathbf{x}}$ in the input space whose image $\phi(\hat{\mathbf{x}})$ is as close as possible to $P_k\phi(\mathbf{x})$. The pre-image problem above introduced thus becomes

$$\hat{\mathbf{x}} = \text{argmin}_{\mathbf{x}\in\mathcal{X}}\|P_k\phi(\mathbf{x}) - \phi(\hat{\mathbf{x}})\|^2_{\mathcal{F}}.$$

In practice, the pre-image problem searches for a map Γ with the property that $\Gamma(\phi(\mathbf{x}_i)) = \mathbf{x}_i$ for $i = 1, \dots, N$. We also require the pre-image map Γ to be decomposed into D (corresponding to the dimension of the input space, i.e. $\dim(\mathcal{X})$) functions so that each component of $\hat{\mathbf{x}}$ is independently estimated. As a result, the proposed method aims to learn a pre-image map constructed as

$$\boldsymbol{\Gamma}(P_k\phi(\mathbf{x})) = \left[\Gamma_1(P_k\phi(\mathbf{x})), \dots, \Gamma_D(P_k\phi(\mathbf{x}))\right]_{1\times D}$$

where the expression for one Γ_j $(j = 1, \dots, D)$ is

$$\Gamma_j(P_k\phi(\mathbf{x})) = \sum_{i=1}^{N}\beta_i^j\ \check{k}\ \left(P_k\phi(\mathbf{x}), P_k\phi(\mathbf{x}_i)\right)$$

and $\check{k}(\cdot,\cdot)$ is a new kernel function which in general may differ from $k(\cdot,\cdot)$. The pre-image problem is therefore reformulated again since each of the D components of $\hat{\mathbf{x}}$ is independently estimated within the input space by employing a new kernel $\check{k}(\cdot,\cdot)$ that projects back the approximated image given by $P_k\phi(\mathbf{x})$. The employed technique to solve such a problem is kernel ridge regression (see for details [46]) which consists of minimising the following function in its dual form below presented:

$$\hat{\boldsymbol{\Gamma}} = \text{argmin}_{\boldsymbol{\Gamma}}\sum_{i=1}^{N} l\left(\mathbf{x}_i - \boldsymbol{\Gamma}(P_k\phi(\mathbf{x}_i))\right) + \lambda R(\boldsymbol{\Gamma})$$

where $\lambda \geq 0$, $R(\boldsymbol{\Gamma})$ is a regularisation term and l is a loss function. To obtain the solution in its dual form let us first define

$$\mathbf{B} = \left[\boldsymbol{\beta}^1, \dots, \boldsymbol{\beta}^D\right] = \begin{bmatrix}\beta_1^1 & \dots & \beta_1^D\\ \vdots & \ddots & \vdots\\ \beta_N^1 & \dots & \beta_N^D\end{bmatrix}_{N\times D}$$

with $\boldsymbol{\beta}^j = (\beta_1^j, \dots, \beta_N^j)^\intercal$ for $j = 1, \dots, D$. By considering the following loss function

$$l\left(\mathbf{x}_i - \boldsymbol{\Gamma}(P_k\phi(\mathbf{x}_i))\right) = \|\mathbf{x}_i - \boldsymbol{\Gamma}(P_k\phi(\mathbf{x}_i))\|^2$$

and the next regularisation form

$$R(\boldsymbol{\Gamma}) = \sum_{j=1}^{N}\|\boldsymbol{\beta}^j\|^2$$

Here we only consider a penalty term on the estimated parameters of the ridge regression, a generalisation would consider a penalisation term also on the hyperparameters of the kernel $\check{k}(\cdot)$, known as the smoothing penalisation term. We consider this more restrictive case since we consider the hyperparameters as known and perform a grid-search introduced in the following section. The criterion exploiting kernel ridge regression can be reformulated in its dual form (see [46] for derivation and details) and the solution is given as

$$\hat{\mathbf{B}}_{N\times D} = \text{argmin}_{\mathbf{B}}\ \text{tr}\left([\mathbf{X} - \check{\mathbf{K}}\mathbf{B}][\mathbf{Y} - \check{\mathbf{K}}\mathbf{B}]^\intercal\right) + \lambda\text{tr}(\mathbf{BB}^\intercal)$$
$$= \left(\check{\mathbf{K}}^\intercal\check{\mathbf{K}} + \lambda\,\mathbf{I}_N\right)^{-1}\check{\mathbf{K}}^\intercal\mathbf{X}$$

where, through the kernel function $\check{k} : \mathbb{R}^Q \times \mathbb{R}^Q \to \mathbb{R}$, we have employed the new Gram Matrix $\check{\mathbf{K}}$ whose entry (i,j) is defined as

$$\check{\mathbf{K}}_{i,j} = \check{k}(P_k\phi(\mathbf{x}_i), P_k\phi(\mathbf{x}_j)) = \check{\phi}(P_k\phi(\mathbf{x}_i))\ \check{\phi}(P_k\phi(\mathbf{x}_j))^\intercal$$

Note that, to introduce the notation for the entire matrix $\check{\mathbf{K}}$, we firstly define $\phi(\mathbf{x}_i)\mathbf{v}_q^\intercal = a_{i,q}$, which add the information of both the ith vector $\mathbf{y}_i$ and the q eigenvector $\mathbf{v}_q$. Hence, the above then becomes

$$\check{\mathbf{K}}_{i,j} = \check{k}(a_{i,q}, a_{j,q}) = \check{\phi}(a_{i,q})\ \check{\phi}(a_{j,q})^\intercal$$

and hence we have

$$bf\,\check{K} = \begin{bmatrix} \check{\phi}_1(a_1) & \dots & \check{\phi}_Q(a_1) \\ \vdots & \ddots & \vdots \\ \check{\phi}_1(a_N) & \dots & \check{\phi}_Q(a_N) \end{bmatrix}_{N\times Q} \begin{bmatrix} \check{\phi}_1(a_1) & \dots & \check{\phi}_1(a_N) \\ \vdots & \ddots & \vdots \\ \check{\phi}_Q(a_1) & \dots & \check{\phi}_Q(a_N) \end{bmatrix}_{Q\times N}$$

Define now the following vector

$$\check{\mathbf{k}}_{\mathbf{x}} = [\check{k}(P_k\phi(\mathbf{y}), P_k\phi(\mathbf{x}_1)), \dots, \check{k}(P_k\phi(\mathbf{x}), P_k\phi(\mathbf{x}_N))]_{1\times N}$$

The pre-image map learns the pre-image $\hat{\mathbf{x}}$ by

$$\hat{\mathbf{x}} = \boldsymbol{\Gamma}(P_k\phi(\mathbf{x})) = (\check{\mathbf{k}}_{\mathbf{x}}\hat{\mathbf{B}})_{1\times D}$$

3.3. CCA for inter data modality cross-correlations

In this section, we review Canonical Correlation Analysis (CCA), its derivation and interpretation, see further details in [47,48]. One of the main advantages in CCA as a multivariate method is that it minimises the risk of committing Type I error, which refers to finding statistically significant results when they do not exist in the population. By allowing for simultaneous comparisons among variables, CCA reduces the need for multiple statistical tests, thereby reducing the experiment-wise error rate. Furthermore, CCA can be used as a comprehensive alternative to other parametric tests commonly used in financial or environmental analysis settings, such as ANOVA, MANOVA, multiple regression, and correlation analysis.

The main appeal of this multivariate method comes from the interest in finding an association between two data sets. If the classical sample correlation matrix is considered, one will obtain the associations between all pairs of variables without having information that allows one to further analyse a decomposition of the within-set associations and the between-set associations (cross-correlations). The objective, when CCA is considered, is to employ a technique that removes the within-set associations to assess the between-set ones and reveal insightful relationships between the two data sets that are hidden and affected by the within-set. Since the objective of CCA is to identify how variations in the data sets can be related, the idea is that each pair of linear combinations must provide distinct pieces of information, achieved by imposing constraints so that each pair of linear combinations of each data set considered are mutually uncorrelated with the other pairs. The correlations between the obtained pairs of linear combinations will be ordered in a decreasing fashion, i.e. the first will carry maximum correlation and the last minimum correlation. The pairs of linear combinations are referred to as canonical functions, where each component of the pairs is referred to as canonical variates. The correlations between the canonical variates are called the canonical correlations. To derive such representations, what is needed is to derive the coefficients of such pairs of linear combinations.

Formally, consider two sets of variables $\mathbf{X} \in \mathbb{R}^{d'}$ and $\mathbf{Y} \in \mathbb{R}^d$, where we assume w.l.o.g. that $d' \leq d$ and we then have

$$\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_{d'} \end{pmatrix}, \quad \mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_d \end{pmatrix}$$

then it is possible to write the full sample correlation matrix as

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{XX} & \boldsymbol{\Sigma}_{XY} \\ \boldsymbol{\Sigma}_{YX} & \boldsymbol{\Sigma}_{YY} \end{pmatrix}$$

where $\boldsymbol{\Sigma}_{XX}$ is the $d' \times d'$ sample correlation matrix of the first sets of variables, $\boldsymbol{\Sigma}_{YY}$ is the $d \times d$ sample correlation matrix of the second sets of variables, $\boldsymbol{\Sigma}_{XY}$ is the $d' \times d$ sample matrix of correlations between the variables of $\mathbf{X}$ and $\mathbf{Y}$, where w.l.o.g. we will assume $d' < d$. $\boldsymbol{\Sigma}_{YX}$ corresponds to the transpose of $\boldsymbol{\Sigma}_{XY}$. CCA corresponds to a parallel method to the PCA applied to the two multivariate data sets (rather than one), looking at linear combinations of paired data. The model proposed by CCA considers two sets of linear combinations, $\mathbf{U} \in \mathbb{R}^{d'}$ and $\mathbf{V} \in \mathbb{R}^{d'}$ respectively, where $\mathbf{U}$ represents the linear combinations of $\boldsymbol{X}$ and $\mathbf{V}$ the linear combinations of $\boldsymbol{Y}$. Each element of U_i is paired with an element of V_i, and these are given as

$$\begin{aligned} U_1 &= a_{11}X_1 + a_{12}X_2 + \cdots + a_{1d'}X_{d'} \\ &\vdots \\ U_{d'} &= a_{d'1}X_1 + a_{d'2}X_2 + \cdots + a_{d'd'}X_{d'} \\ V_1 &= b_{11}Y_1 + b_{12}Y_2 + \cdots + b_{1d}Y_d \\ &\vdots \\ V_{d'} &= b_{d'1}Y_1 + b_{d'2}Y_2 + \cdots + b_{d'd}Y_d \end{aligned}$$

where, each ith pair (U_i, V_i) corresponds to the canonical variate. The extraction of the canonical variables proceeds as a sequence of increasingly constrained optimisations which can be formally expressed as follows. Given column random vectors $\mathbf{X}_i \in \mathbb{R}^{d'}$ and $\mathbf{Y}_i \in \mathbb{R}^d$ with finite second moments, with $\min\{d', d\}$ variates extracted, CCA seeks vectors $\boldsymbol{a} \in \mathbb{R}^{d'}$ and $\boldsymbol{b} \in \mathbb{R}^d$ such that the random variables $\boldsymbol{a}^T\boldsymbol{X}_i$ and $\boldsymbol{b}^T\boldsymbol{Y}_i$ maximise correlation

$$\begin{aligned} \boldsymbol{a}_i, \boldsymbol{b}_i &= \arg\max_{\boldsymbol{a},\boldsymbol{b}} Corr\left(U_i, V_i\right) \\ &= \arg\max_{\boldsymbol{a},\boldsymbol{b}} \frac{Cov\left(\boldsymbol{a}^T\boldsymbol{X}_i, \boldsymbol{b}^T\boldsymbol{Y}_i\right)}{\sqrt{Var\left(\boldsymbol{a}^T\boldsymbol{X}_i\right) Var\left(\boldsymbol{b}^T\boldsymbol{Y}_i\right)}} \end{aligned} \tag{5}$$

subject to constraint set

$$\begin{aligned} &Var\left(U_i\right) = Var\left(V_i\right) = 1 \\ &\left\{Corr\left(U_j, U_i\right) = Corr\left(V_j, V_i\right) = 0\right\}_{j=1}^{i-1} \\ &\left\{Corr\left(U_j, V_i\right) = Corr\left(U_i, V_j\right) = 0\right\}_{j=1}^{i-1} \end{aligned}$$

The solution to this sequence of constrained optimisations can then be expressed in a matrix form where the optimal coefficients $\{\mathbf{a}_i\}_{i=1}^{d'}$ can be show to be the eigenvectors of the matrix

$$\boldsymbol{\Sigma}_{XX}^{-1}\boldsymbol{\Sigma}_{XY}\boldsymbol{\Sigma}_{YY}^{-1}\boldsymbol{\Sigma}_{YX}$$

and $\{\mathbf{b}_i\}_{i=1}^{d'}$ are the eigenvectors of the matrix

$$\boldsymbol{\Sigma}_{YY}^{-1}\boldsymbol{\Sigma}_{YX}\boldsymbol{\Sigma}_{XX}^{-1}\boldsymbol{\Sigma}_{XY}$$

Furthermore, the canonical correlation will correspond to the square roots of the non-zero eigenvalues of the above matrices.

From this discussion it is clear that CCA examines the correlation between synthetic variables (canonical variates) weighted according to the relationships between the original variables. It can be seen as a simple bivariate correlation between the two artificially constructed variables, i.e. (U_i, V_i). The information captured by the canonical correlation $\rho_1^\star, \rho_2^\star, \dots, \rho_{d'}^\star$ (since the maximum number of canonical correlations is the minimum number of variables of the two data sets considered) represents the associations between the set of $\mathbf{X}$ and the set of $\mathbf{Y}$ after the within-set correlations have been removed.

Having extracted the optimal canonical correlations denoted by $\{\rho_i^*\}_{i=1}^{d'}$ that seek to estimate the true population values denoted $\{\rho_i\}_{i=1}^{d'}$, one may wish to assess the statistical significance of the canonical correlations under a hypothesis test in which the null hypothesis is stated as follows:

$H_0 : \rho_1 = \rho_2 = \cdots = \rho_{d'} = 0$

H_1 : At least one canonical correlation significantly differs from zero

A test of this hypotheses can be applied, given the null, using a test statistic known as Wilks' lambda: $\Lambda = \prod_{i=1}^{d'}(1-\rho_i^{*2})$, where ρ_i^* is the estimated optimal canonical correlation for the ith pair (u_i, v_i) [49]. Bartlett showed that under the null hypothesis, a particular function of Λ would be distributed approximately as a chi-squared variate [49]. Therefore the statistical significance of Wilk's Λ requires the calculation of the following statistic:

$$\chi^2 = -[N - 0.5(d + d' + 1)] \log \Lambda$$

where N is the number of samples and d and d' are the number of variables in X and Y respectively. For this test, if H_0 is rejected, then Bartlett proposed [49] a sequence of procedures that test whether the second-largest canonical correlation significantly differs from zero, then the third largest, etc. Of the methods proposed by Bartlett in this context, the most used in practice, given its robustness to smaller sample sizes than the χ^2 test, is the one following Rao's F-approximation [50]. This employs a likelihood ratio test approach, see details in [51]. This test proceeds as follows, starting with $i = 0$, the null hypothesis tests

$$H_0 : \check{d} = i$$
$$H_1 : \check{d} > i$$

where $\check{d}$ is related to the order spline basis model representation used for each dataset when constructing the CCA (see [51], Eq. (8) and following discussion for further details). If H_0 is rejected, i is incremented and a new test is conducted. This proceeds while H_0 is not rejected or i reaches $M = min(\check{d}, \check{d}')$. We can see the relationship between Wilk's Λ test statistic and Rao's F-approximation test statistic as follows for a given number of m canonical variates:

$$F_{\mathrm{df}_1,\mathrm{df}_2} = \frac{\mathrm{df}_2}{\mathrm{df}_1}\left(\frac{1-\Lambda_m}{\Lambda_m}\right)^{1/\nu}$$

where $\nu = \sqrt{(\mathrm{df}_1^2 - 4)/((d'-m)^2 + (d-m)^2 - 5)}$, $\mathrm{df}_1 = (d'-m)(d-m)$, $\mathrm{df}_2 = (N - 1.5 - (d'+d)/2)\nu - \mathrm{df}_1/2 + 1$ (where N is the number of samples) and Wilk's test statistics is given by

$$\Lambda_m = \prod_{i=m+1}^{d'}(1-\rho_i^{*2})$$

This Rao's F-approximation test is employed in the results analysis.

What remains is to discuss how to interpret the canonical variates that are found to be statistically significant. A summary of relevant quantities one can obtain from CCA analysis is provided in Table 1 and described further below. Firstly, *the canonical correlation coefficient* quantifies the strength of the association between the two sets of variables. It represents the maximum correlation achievable between linear combinations of variables from the two sets. It ranges from 0 to 1, with 0 indicating no relationship and 1 indicating a perfect linear relationship. It is similar to the multiple R value used in regression analysis. Secondly, *the squared canonical correlations* represent the proportion of shared variance between the synthetic variables in each canonical function. They indicate how much of the variance in one set of variables can be explained by the other set. Thirdly, another quantity often examined is the *redundancy index* which corresponds to a measure of the total amount of variance explained in a set of variables by all the combined canonical functions. It represents the cumulative proportion of variance taken into account in the original set of variables. In other words, it quantifies the overall redundancy or overlap between the two sets of variables. One could compare this index to the factor loadings in factor analysis which represent the proportion of variance in each observed variable accounted for by the underlying latent factors. Alternatively, in structural equation modelling, the squared multiple correlations (R^2) indicate the amount of variance in an observed variable explained by its associated latent variable. Although squared canonical correlations focus on the specific relationship between individual canonical functions and their associated synthetic variables, the redundancy index provides a broader summary of the overall explanatory power of all combined canonical functions. High redundancy suggests a high ability to predict. Fourthly, the *canonical function* can be thought of as a derived synthetic variable that represents the relationship or association between the two sets of original variables. Each function is orthogonal to every other function, properties that make them analogous to components in a principal component analysis. Furthermore, this orthogonality allows one to interpret each function separately. A single function can be comparable to the set of standardised weights found in multiple regression (albeit only for the predictor variables). *Standardised canonical function coefficients* refer to coefficients that have been standardised and are used in linear combinations to merge observed variables into two synthetic variables. These weights are applied to the observed scores in Z-score form to generate the synthetic scores, which are then correlated to determine the canonical correlation. These coefficients are derived to maximise this canonical correlation and can be directly compared to beta weights in regression analysis. A *structure coefficient* is the bivariate correlation between an observed variable and a synthetic variable. Since these coefficients are Pearson R statistics, they may range from -1 to $+1$. In practice, they provide information about which of the original variables are useful in defining the synthetic ones, i.e. the canonical variate, within the CCA model. Such coefficients are analogous to the structure coefficients of the matrix of factor analysis structure or in a multiple regression as the correlation between a predictor and the predicted Y' scores [52,53]. Lastly, the *Squared canonical structure coefficients* are the square of the structure coefficients. This statistic is analogous to any other R^2-type effect size and indicates the proportion of variance an observed variable linearly shares with the synthetic variable generated from the observed variable's set.

Robust alternatives to CCA has been proposed in the literature [54]. Such robust methods are designed to handle non-normal or non-linear data. In this vain, the idea of this work is to propose a method combines non-linear feature extraction per data set, achieved by first applying the kPCA and extracting a set of kPCs and then to apply the CCA to these factors to capture the cross-correlation of these basis functions. To better understand this point, we derive a worked comparison between a linear method based on PCA-CCA combination versus how the kPCA-CCA method proposed differs, see Section 4.1. The purpose of this worked explicit derivation is to explicitly show the mathematical steps involved in such an approach in the case of a linear kernel which corresponds to a PCA-CCA method then we show what is modified in this example when a more general non-linear kernel is considered.

In the following discussion in Section 4 the core of the methodological contributions are presented which combine the ideas of kPCA feature extraction per data set with the utilisation of CCA methods to assess associations of relevance between features from each data set.

4. Analysis of multi-modal spatial-temporal factors from multiple data sets via kPCA-CCA

We develop a novel solution to detect cross-correlation between spatial–temporal multivariate data sets, which will accommodate to the presence of non-stationary and non-linear spatial–temporal data generating processes as well as the presence of different types of structured data, i.e. numerical or categorical or different timestamps observation frequencies and spatial locations. The procedure is a two stage process: firstly exploiting the previously presented ideas of kPCA to extract non-stationary and non-linear basis components for each of the data sets considered. Secondly, we then apply CCA methods to study relevant statistical associations between the factors extracted from the datasets via the kPCA method. Several spatial–temporal multivariate data sets will be analysed, each describing pollution (or climate) conditions for every county in California, along with another data set detailing municipal green bonds for each county of California. The primary goal is to comprehensively characterise the variations present in these multi-multivariate data sets simultaneously, treating the fluctuations of

Table 1
Quantities required for assessing the CCA model.

CCA Model Assessment

Quantity	Notation	Interpretation	Relation with other models
Canonical Correlation Coeff.	ρ_i^*	Association between **X** and **Y**	Similar to the multiple R value in regression
Squared Canonical Correlation	ρ_i^{*2}	Proportion of shared variance between the synthetic variables in each canonical function	Analogous to the R^2 effect in multiple regression
Redundancy Index	$(\sum_{j=1}^{d'} Corr^2\left(Y_j, V_i\right)/d')\rho_j^{*2}$	Cumulative proportion of variance accounted for in the original variable set	Analogous to the factor loadings in factor analysis
Canonical Variates	U_i, V_i	Individual variable of the synthetic pairs	Analogous to PCA bases or factor scores in factor analysis
Canonical Function	(U_i, V_i)	Synthetic variable pair for the association between **X** and **Y**	Analogous to PCA bases or factor scores in factor analysis
Canonical Coefficients	$\mathbf{a}_i, \mathbf{b}_i$	Coeffs.maximising the canonical correlation	Equivalent to beta weights in regression
Structure Coeff. (Canonical Loadings)	$Corr\left(X_i, U_j\right)$, $Corr\left(Y_i, V_j\right)$	Bivariate correlation between an observed variable and a synthetic one	Analogous to the correlation between a predictor and the predicted Y' scores in a multiple regression
Squared Canonical Structure Coeff.	$Corr^2\left(X_i, U_j\right)$ $Corr^2\left(Y_i, V_j\right)$	Proportion of variance an observed variable linearly shares with the synthetic variable	Analogous to any other R^2-type effect size

municipal green bonds in the financial market as a collective entity, and the variations in pollution (and climate) across California as a global change. By adopting this methodology, cross-correlations of these data, between different counties, will be quantified, providing insights into the inter-dependencies and relationships across the counties. The procedure will therefore apply kPCA to each data set of every county and will obtain a set of kPCs for financial data for Alameda, Los Angeles, Napa, etc., and a set of kPCs for pollution and climate data at a county level. Once the most significant kPCs are retained, they will be fed to the CCA to observe cross-correlation between the modes carrying maximum variation across counties. This will be done one kPC per time, i.e. the kPC1 of the first data set for all the counties vs the kPC1 of the second data set for all the corresponding counties, etc. In such a way, every variation mode will be related to the ones of other data sets, and the presence of correlation as well as its direction will be interpreted. The method will be known as kPCA-CCA. Its benchmark comparison will be its linear version, i.e. PCA-CCA, which will be constructed equivalently, but PCs will be used instead, analogous to the simplest kPCA using a linear kernel.

In developing kPCA-CCA, we note that whilst the individual three data sets each have a variety of different multivariate spatial–temporal time series features included in each of the pollution, climate and green bond data sets, they do not need to be reported at the same time intervals or on a regular set of time intervals. This is one of the additional advantages of using the kPCA methodology to extract features from each data set, since after feature extraction under kPCA the resulting kPCs can all then be evaluated out-of-sample at a new fixed common set of feature points, i.e. on a new mesh, common across the counties, as per Eq. (4). By exploiting the out-of-sample problem presented in Section 3.2.4, we will obtain a new set of kPCs which are comparable on a common spatial–temporal grid. Fig. 1 summarises this implemented methodology.

Remark that a parametric kernel function incorporates one or more hyperparameters, which partly control the underlying feature samples similarity structure. One of the main issues when using kernel methods is indeed the learning of such hyperparameters. Our procedure exploits a grid search over a set of pre-chosen hyperparameters and then projects back the mapped data points through the pre-image method. The set providing the minimum euclidean distance to the original data points will be the selected set of hyperparameters. We offered a detailed explanation of such a procedure in Section 3.2.5. Furthermore, the considered financial data set contains multiple sources of data, i.e. categorical, numerical, etc. and this will cause issues for the kPCs extraction since a unique kernel function must be computed. We offer a method that is able to deal with such different data sources and relies on the Jaccard distance and Jaccard kernel below introduced.

Key details regarding practical implementation features of the method described relating to the selection and construction of the common spatial–temporal mesh used to standardise the representations of the kPCA across the different data sets, the details of the hyper parameter learning method developed for the kernel parameters optimal selection are provided in detail in the Supplementary Information accompanying this manuscript.

4.1. Comparing CCA, PCA-CCA and kPCA-CCA

In this subsection we explicitly compare the benchmark reference linear projection method of PCA-CCA versus the proposed more general non-linear method of kPCA-CCA. The idea is to mathematically explain the framework proposed in Fig. 1. In so doing, we seek to provide a straightforward interpretation of why we combine the CCA in conjunction with the method of kPCA.

To illustrate the advantages of using kPCA-CCA over CCA or PCA-CCA, we consider an example involving two datasets. By applying each of these methodologies, we observe how the yielded canonical correlation coefficients measure different quantities. We begin by examining the standard CCA approach without prior projections, where the derived canonical correlation coefficients detect correlation between linear combinations of the original data sets.

This two-stage process leads to distinctive formulations for CCA, with key insights derived from the eigen decomposition of covariance matrices. Notably, PCA-CCA provides canonical correlations that are weighted by the inverse of the eigenvalues of the covariance matrices of the PCA-transformed data sets. These eigenvalues represent the variance of the data along the principal component axes, emphasising the contributions of the principal components with higher variances. This weighting scheme helps in identifying the most significant relationships between the two data sets, reflected in the canonical correlations.

Following the PCA-CCA exposition, we transition to the kPCA-CCA methodology, which integrates non-linear kernel PCA into the CCA framework. Here, each dataset undergoes non-linear transformation via kernel PCA before being subjected to CCA. Crucially, kPCA-CCA enables capturing non-linear relationships between variables, offering enhanced flexibility over traditional linear methods. The derivation shows that canonical correlations in kPCA-CCA are weighted by the inverse of the eigenvalues of the kernel matrices, which represent similarities or distances between data points in the kernel-induced feature space.

In this section we derive the differences in canonical correlation coefficients among the three methods, which stem from the different ways the data are transformed When applying CCA. CCA directly assesses correlations between original variables, PCA-CCA focuses on linear transformations capturing linearly maximum variance and kPCA-CCA captures non-linear maximum variance relationships. Each method's choice impacts how the relationships between variables are modelled and, consequently, the resulting canonical correlation coefficients.

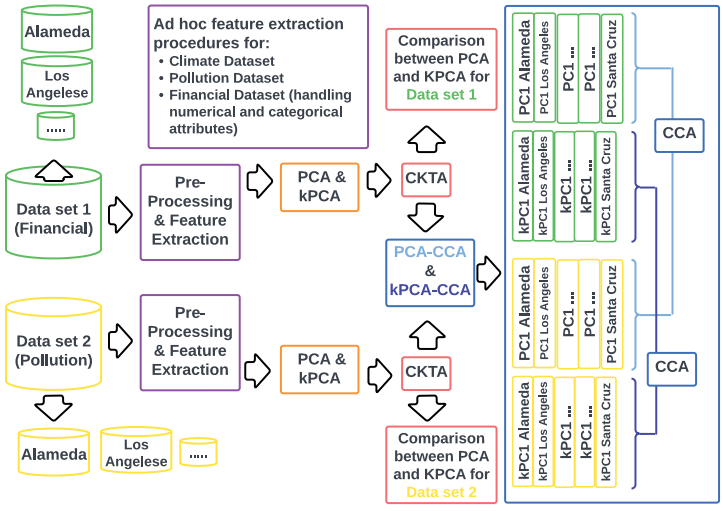

Fig. 1. Stages of kPCA-CCA methodology. The analysis will be performed by running kPCA-CCA (and PCA-CCA benchmark comparison) on pollution vs financial kPCs and climate vs financial kPCs. Given the two data sets, say pollution and financial data set, we split them by county and perform a pre-cleaning and pre-processing procedures as required to remove data issues such as NA's and misreporting. Since each data set has multiple attributes observed at multiple monitoring sites in each county, some work is first performed to combine and clean the data in the Pre-processing stage (details are available in the associated github repository and the tools description for these stages, see ....). Next, linear PCA and the non-linear kPCA are applied per dataset and, for each decomposition method, the first three bases (PC1, PC2, PC3) or (kPC1, kPC2, kPC3) are retained. The next step consists of analysis of which extracted features better capture the variability of the given data set via the method of centered kernel target alignment (cKTA), which will be introduced in subsections below. Finally, the PCA-CCA and the kPCA-CCA will be computed between pollution and financial PCs and kPCs, respectively, to measure the impact of green bonds on pollution attributes within different counties in California. Note that the CCA is run by considering all counties for individual group of bases, i.e. kPC1 financial vs kPC1 pollution, kPC2 financial vs kPC2 pollution, etc. and PC1 financial vs PC1 pollution, PC2 financial vs PC2 pollution, and so on.

4.1.1. Comparative analysis between different projection methods with CCA: Illustrative closed form example

Consider two multivariate data sets $\mathbf{X}_{N\times d'}$ and $\mathbf{Y}_{N\times d}$, where each data point $\mathbf{X}_i$ and $\mathbf{Y}_i$ is assumed to be normally distributed. Specifically, the joint distribution of $\mathbf{X}_i$ and $\mathbf{Y}_i$ is given by a multivariate normal distribution $\mathcal{N}$ given as

$$\begin{pmatrix}\mathbf{X}_i\\ \mathbf{Y}_i\end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix}\boldsymbol{\mu}\\ \boldsymbol{\nu}\end{pmatrix}, \begin{pmatrix}\boldsymbol{\Sigma}_{XX} & \boldsymbol{\Sigma}_{XY}\\ \boldsymbol{\Sigma}_{YX} & \boldsymbol{\Sigma}_{YY}\end{pmatrix}\right)$$

where $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$ represent the mean vectors of $\mathbf{X}_i$ and $\mathbf{Y}_i$ respectively, while $\boldsymbol{\Sigma}_{XX}$, $\boldsymbol{\Sigma}_{YY}$, and $\boldsymbol{\Sigma}_{XY} = \boldsymbol{\Sigma}_{YX}^{\top}$ denote the covariance matrices of $\mathbf{X}_i$, $\mathbf{Y}_i$, and the cross-covariance matrix between $\mathbf{X}_i$ and $\mathbf{Y}_i$ and can be formally introduced as

$$\begin{aligned}\mathbb{V}ar(\mathbf{X}_i) &= \boldsymbol{\Sigma}_{XX_{(d'\times d')}}\\ \mathbb{V}ar(\mathbf{Y}_i) &= \boldsymbol{\Sigma}_{YY_{(d\times d)}}\\ \mathbb{C}ov(\mathbf{X}_i,\mathbf{Y}_i) &= \boldsymbol{\Sigma}_{XY_{(d'\times P)}} = \boldsymbol{\Sigma}_{YX_{(d'\times d)}}\end{aligned}$$

The objective is to detect the spatio-temporal cross-correlation between $\mathbf{X}_{N\times d'}$ and $\mathbf{Y}_{N\times d}$, capturing the statistical association between the two data sets over time and space. This analysis involves identifying the relationships between variables in $\mathbf{X}_{N\times d'}$ and $\mathbf{Y}_{N\times d}$, quantified by canonical correlation coefficients, providing valuable insights into the underlying relationships and dynamics of the data sets. By integrating kPCA and CCA, our approach can effectively model the intricate spatio-temporal dependencies present in $\mathbf{X}_{N\times d'}$ and $\mathbf{Y}_{N\times d}$. In an ideal scenario with independent and identically distributed (i.i.d.) multivariate Gaussian observations, the relationships between CCA, PCA-CCA and kPCA-CCA are straightforward to discern. These methods perform similar tasks in this ideal setting, albeit with different approaches. CCA seeks linear projections maximising the correlation between two datasets, PCA-CCA first conducts PCA for dimensionality reduction before CCA, and kPCA-CCA utilises kernel PCA to handle nonlinearity before CCA. However, the distinctions between these methods become more pronounced in real-world settings characterised by complexity, such as non-Gaussian distributions, nonlinear relationships, and various noise sources. In such contexts, kPCA-CCA is particularly valuable due to its ability to capture nonlinear relationships through the kernel trick. This enables it to uncover latent patterns and structures that might elude linear methods like PCA-CCA and CCA. Therefore, while the direct relationships between these methods are evident in ideal scenarios, kPCA-CCA's differentiation becomes more pronounced and advantageous in navigating the complexities of real-world data analysis. In the following subsections, we derive CCA, PCA-CCA and kPCA-CCA solutions for this example and show how the obtained correlation coefficients differ from each case and the ones computed over the kPCA incorporate non-linear solutions.

4.1.2. Only applying CCA method

If one first applied the CCA method, explained in Section 3.3, directly to the two data sets without any prior projections. Then the correlation between any linear projections of the two data sets is given by

$$\rho(\mathbf{a},\mathbf{b}) = \frac{\mathbf{a}^{\intercal}\Sigma_{XY}\mathbf{b}}{(\mathbf{a}^{\intercal}\Sigma_{XX}\mathbf{a})^{1/2}(\mathbf{b}^{\intercal}\Sigma_{YY}\mathbf{b})^{1/2}}$$

Note that the property of scale invariance applies as follows

$$\rho(c\,\mathbf{a},\mathbf{b}) = c\,\rho(\mathbf{a},\mathbf{b})$$

Therefore, in this case, the CCA problem can be expressed as the solution to the optimisation given by

$$\begin{aligned}&\max_{\mathbf{a},\mathbf{b}}\ \mathbf{a}^{\intercal}\boldsymbol{\Sigma}_{XY}\mathbf{b}\\ \text{s.t.}\quad &\mathbf{a}^{\intercal}\boldsymbol{\Sigma}_{XX}\mathbf{a} = 1\\ &\mathbf{b}^{\intercal}\boldsymbol{\Sigma}_{YY}\mathbf{b} = 1.\end{aligned}$$

In the following, one can show how to solve this CCA optimisation objective by first defining the matrix

$$\boldsymbol{\Gamma} = \boldsymbol{\Sigma}_{XX}^{-1/2}\ \boldsymbol{\Sigma}_{XY}\ \boldsymbol{\Sigma}_{YY}^{-1/2}$$

and then applying the Singular Value Decomposition (SVD) to obtain

$$\boldsymbol{\Gamma} = \mathbf{W}\,\mathbf{D}\,\mathbf{V}^{\intercal}$$

where

$$\mathbf{W} = (\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_k)$$
$$\mathbf{D} = \text{diag}(\sqrt{\lambda_1}, \sqrt{\lambda_2}, \dots, \sqrt{\lambda_k})$$
$$\mathbf{V} = (\boldsymbol{\nu}_1, \boldsymbol{\nu}_2, \dots, \boldsymbol{\nu}_k)$$

where

$$k = \text{rank}(\boldsymbol{\Gamma}) = \text{rank}(\boldsymbol{\Sigma}_{XY}) = \text{rank}(\boldsymbol{\Sigma}_{YX})$$

with $\lambda_1 > \lambda_2 > \lambda_3 > \cdots > \lambda_k$ are non-zero eigenvalues of the matrix $\boldsymbol{\Gamma}\boldsymbol{\Gamma}^\intercal_{(d'\times d')}$ or $\boldsymbol{\Gamma}^\intercal\boldsymbol{\Gamma}_{(d\times d)}$. Note that $\mathbf{w}_i$ and $\boldsymbol{\nu}_i$ are the standardised eigenvectors of $\boldsymbol{\Gamma}\boldsymbol{\Gamma}^\intercal_{(d'\times d')}$ and $\boldsymbol{\Gamma}^\intercal\boldsymbol{\Gamma}_{(d\times d)}$ respectively.

From these matrices, one can then find the canonical coefficients for $i \in 1, \dots, k$ as follows

$$\mathbf{a}_i = \boldsymbol{\Sigma}_{XX}^{-1/2}\mathbf{w}_i$$
$$\mathbf{b}_i = \boldsymbol{\Sigma}_{YY}^{-1/2}\boldsymbol{\nu}_i$$

Applying these canonical coefficients to project columns of each data set, one obtains the canonical correlation variables (U_i, V_i)

$$\eta_i = \mathbf{a}_i^\intercal\mathbf{X}_i$$
$$\psi_i = \mathbf{b}_i^\intercal\mathbf{Y}_i$$

From these we can conclude that the canonical correlation coefficients measure correlation between linear combinations in each group of original variables $\mathbf{X}_{N\times d'}$ and $\mathbf{Y}_{N\times d}$. Note that the squared coefficients correspond to the eigenvalues or canonical roots of the square matrices

$$\underbrace{\boldsymbol{\Gamma}\boldsymbol{\Gamma}^\intercal}_{d'\times d'} = \underbrace{(\mathbf{X}^\intercal\mathbf{X})^{-1}}_{\boldsymbol{\Sigma}_{XX}^{-1}}\underbrace{\mathbf{X}^\intercal\mathbf{Y}}_{\boldsymbol{\Sigma}_{XY}}\underbrace{(\mathbf{Y}^\intercal\mathbf{Y})^{-1}}_{\boldsymbol{\Sigma}_{YY}^{-1}}\underbrace{\mathbf{Y}^\intercal\mathbf{X}}_{\boldsymbol{\Sigma}_{YX}}$$
$$\underbrace{\boldsymbol{\Gamma}^\intercal\boldsymbol{\Gamma}}_{d\times d} = (\mathbf{Y}^\intercal\mathbf{Y})^{-1}\mathbf{Y}^\intercal\mathbf{X}(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal\mathbf{Y}$$

The first eigenvalue accounts for the highest correlation between the pairs of canonical variates and the rest of the eigenvalues are obtained in descending order of correlation. Furthermore, the coefficients defining the canonical variates are obtained as eigenvectors associated to the highest canonical roots in the square matrices, i.e. the first eigenvalue. The coefficients for vector $\mathbf{a}$ are in $\boldsymbol{\Gamma}\boldsymbol{\Gamma}^\intercal$ while the coefficients for vector $\mathbf{b}$ are in $\boldsymbol{\Gamma}^\intercal\boldsymbol{\Gamma}$.

4.1.3. Applying PCA followed by CCA method (PCA-CCA)

Next, consider what happens if we first take the PCA projection for each data set $\mathbf{X}_{N\times d'}$ and $\mathbf{Y}_{N\times d}$ individually and then repeat the application of CCA on the transformed PCA projected data sets that will be denoted generically by $\widetilde{\mathbf{X}}_{N\times p'} = \widetilde{\mathbf{X}}_{N\times d'}\mathbf{W}_{\mathbf{1}d'\times p'}$ and $\widetilde{\mathbf{Y}}_{N\times p} = \widetilde{\mathbf{Y}}_{N\times d}\mathbf{W}_{\mathbf{2}d\times p}$ for $p \le d$ and $p' \le d'$. Under this two stage process, in the second stage of the CCA, of the PCA transformed variables, we will have a formulation given as follows

$$\tilde{\mathbf{X}} = \mathbf{X}\mathbf{W}_1 \quad \text{for } \mathbf{W}_1 \text{ PCs s.t.}$$
$$\mathbf{W}_1^\intercal\mathbf{W}_1 = \mathbb{I}_{d'}$$
$$\boldsymbol{\Sigma}_{XX}\mathbf{W}_1 = \boldsymbol{\Lambda}_1\mathbf{W}_1$$
$$\tilde{\mathbf{X}}_i, \tilde{\mathbf{X}}_j \text{ are independent}$$

and

$$\tilde{\mathbf{Y}} = \mathbf{Y}\mathbf{W}_2 \quad \text{for } \mathbf{W}_2 \text{ PCs s.t.}$$
$$\mathbf{W}_2^\intercal\mathbf{W}_2 = \mathbb{I}_P$$
$$\boldsymbol{\Sigma}_{YY}\mathbf{W}_2 = \boldsymbol{\Lambda}_2\mathbf{W}_2$$
$$\tilde{\mathbf{Y}}_i, \tilde{\mathbf{Y}}_j \text{ are independent}$$

Suppose we retain all the PCs for both $\mathbf{X}$ and $\mathbf{Y}$ i.e. $p' = d'$ and $p = d$. Then according to the formulation in Section 4.1.2 the CCA is obtained from the matrices

$$\tilde{\boldsymbol{\Gamma}}\tilde{\boldsymbol{\Gamma}}^\intercal = (\tilde{\mathbf{X}}^\intercal\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}^\intercal\tilde{\mathbf{Y}}(\tilde{\mathbf{Y}}^\intercal\tilde{\mathbf{Y}})^{-1}\tilde{\mathbf{Y}}^\intercal\tilde{\mathbf{X}}$$
$$\tilde{\boldsymbol{\Gamma}}^\intercal\tilde{\boldsymbol{\Gamma}} = (\tilde{\mathbf{Y}}^\intercal\tilde{\mathbf{Y}})^{-1}\tilde{\mathbf{Y}}^\intercal\tilde{\mathbf{X}}(\tilde{\mathbf{X}}^\intercal\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}^\intercal\tilde{\mathbf{Y}}$$

Note that the columns of $\tilde{\mathbf{X}}$ and $\tilde{\mathbf{Y}}$ are orthonormal after the PCA projections. Substituting the transformation and rearranging the algebra gives

$$\underbrace{\tilde{\boldsymbol{\Gamma}}\tilde{\boldsymbol{\Gamma}}^\intercal}_{d'\times d'} = ((\mathbf{X}\mathbf{W}_1)^\intercal(\mathbf{X}\mathbf{W}_1))^{-1}(\mathbf{X}\mathbf{W}_1)^\intercal(\mathbf{Y}\mathbf{W}_2)((\mathbf{Y}\mathbf{W}_2)^\intercal(\mathbf{Y}\mathbf{W}_2))^{-1}(\mathbf{Y}\mathbf{W}_2)^\intercal(\mathbf{X}\mathbf{W}_1)$$
$$\underbrace{\tilde{\boldsymbol{\Gamma}}^\intercal\tilde{\boldsymbol{\Gamma}}}_{d\times d} = ((\mathbf{Y}\mathbf{W}_2)^\intercal(\mathbf{Y}\mathbf{W}_2))^{-1}(\mathbf{Y}\mathbf{W}_2)^\intercal(\mathbf{X}\mathbf{W}_1)((\mathbf{X}\mathbf{W}_1)^\intercal(\mathbf{X}\mathbf{W}_1))^{-1}(\mathbf{X}\mathbf{W}_1)^\intercal(\mathbf{Y}\mathbf{W}_2)$$

Then, using the fact that $\mathbf{W}_1$ and $\mathbf{W}_2$ correspond to the PCA projections for the two data set respectively, one obtains

$$\underbrace{\tilde{\boldsymbol{\Gamma}}\tilde{\boldsymbol{\Gamma}}^\intercal}_{d'\times d'} = \underbrace{\boldsymbol{\Lambda}_1^{-1}}_{d'\times d'}\underbrace{(\mathbf{X}\mathbf{W}_1)^\intercal}_{d'\times N}\underbrace{(\mathbf{Y}\mathbf{W}_2)}_{N\times d}\underbrace{\boldsymbol{\Lambda}_2^{-1}}_{d\times d}\underbrace{(\mathbf{Y}\mathbf{W}_2)^\intercal}_{d\times N}\underbrace{\mathbf{X}\mathbf{W}_1}_{N\times d'}$$
$$\underbrace{\tilde{\boldsymbol{\Gamma}}^\intercal\tilde{\boldsymbol{\Gamma}}}_{d\times d} = \underbrace{\boldsymbol{\Lambda}_2^{-1}}_{d\times d}\underbrace{(\mathbf{Y}\mathbf{W}_2)^\intercal}_{d\times N}\underbrace{(\mathbf{X}\mathbf{W}_1)}_{N\times d'}\underbrace{\boldsymbol{\Lambda}_1^{-1}}_{d'\times d'}\underbrace{(\mathbf{X}\mathbf{W}_1)^\intercal}_{d'\times N}\underbrace{(\mathbf{Y}\mathbf{W}_2)}_{N\times d}$$

If $d = d'$, then

$$\begin{aligned}\tilde{\boldsymbol{\Gamma}}\tilde{\boldsymbol{\Gamma}}^\intercal &= \boldsymbol{\Lambda}_1^{-1}\boldsymbol{\Lambda}_2^{-1}(\mathbf{X}\mathbf{W}_1)^\intercal(\mathbf{Y}\mathbf{W}_2)(\mathbf{W}_2^\intercal\mathbf{Y}^\intercal)(\mathbf{X}\mathbf{W}_1)\\ &= \boldsymbol{\Lambda}_1^{-1}\boldsymbol{\Lambda}_2^{-1}(\mathbf{X}\mathbf{W}_1)^\intercal(\mathbf{Y}\mathbf{Y}^\intercal)(\mathbf{X}\mathbf{W}_1)\\ &= \boldsymbol{\Lambda}_1^{-1}\boldsymbol{\Lambda}_2^{-1}(\mathbf{Y}\mathbf{Y}^\intercal)(\mathbf{W}_1^\intercal\mathbf{X}^\intercal)(\mathbf{X}\mathbf{W}_1)\\ &= \boldsymbol{\Lambda}_1^{-1}\boldsymbol{\Lambda}_2^{-1}(\mathbf{Y}\mathbf{Y}^\intercal)\boldsymbol{\Lambda}_1\\ &= \boldsymbol{\Lambda}_2^{-1}(\mathbf{Y}\mathbf{Y}^\intercal)\end{aligned}$$

and

$$\begin{aligned}\tilde{\boldsymbol{\Gamma}}^\intercal\tilde{\boldsymbol{\Gamma}} &= \boldsymbol{\Lambda}_2^{-1}\boldsymbol{\Lambda}_1^{-1}(\mathbf{Y}\mathbf{W}_2)^\intercal(\mathbf{X}\mathbf{W}_1)(\mathbf{W}_1^\intercal\mathbf{X}^\intercal)(\mathbf{Y}\mathbf{W}_2)\\ &= \boldsymbol{\Lambda}_2^{-1}\boldsymbol{\Lambda}_1^{-1}(\mathbf{Y}\mathbf{W}_2)^\intercal(\mathbf{X}\mathbf{X}^\intercal)(\mathbf{Y}\mathbf{W}_2)\\ &= \boldsymbol{\Lambda}_2^{-1}\boldsymbol{\Lambda}_1^{-1}(\mathbf{X}\mathbf{X}^\intercal)(\mathbf{W}_2^\intercal\mathbf{Y}^\intercal)(\mathbf{Y}\mathbf{W}_2)\\ &= \boldsymbol{\Lambda}_2^{-1}\boldsymbol{\Lambda}_1^{-1}(\mathbf{X}\mathbf{X}^\intercal)\boldsymbol{\Lambda}_2\\ &= \boldsymbol{\Lambda}_1^{-1}(\mathbf{X}\mathbf{X}^\intercal)\end{aligned}$$

Hence, in the case of the PCA-CCA, the coefficients $\mathbf{a}_i$ and $\mathbf{b}_i$ will be the eigenvectors associated to the highest canonical roots in the matrices $\boldsymbol{\Lambda}_2^{-1}(\mathbf{Y}\mathbf{Y}^\intercal)$ and $\boldsymbol{\Lambda}_1^{-1}(\mathbf{X}\mathbf{X}^\intercal)$ respectively. Next we see what differs from the PCA-CCA linear method when compared to the non-linear functional kPCA version of this two stage procedure, that we call kPCA-CCA.

4.1.4. Applying kPCA followed by CCA method (kPCA-CCA)

In the kernel version of this hybrid method, denoted by kPCA-CCA, as before we first transform each data set, this time using non-linear kernel kPCA method for each data set, giving

$$\underbrace{\mathbf{A}_1}_{N\times p'} = \underbrace{\mathbf{K}_1}_{N\times N}\underbrace{\mathbf{W}_1}_{N\times p'} \quad \text{for } \mathbf{W}_1 \text{ kPCs s.t.}$$
$$\mathbf{W}_1^\intercal\mathbf{W}_1 = \mathbb{I}_{p'}$$
$$\mathbf{W}_1^\intercal\mathbf{K}_1 = \boldsymbol{\Lambda}_1\mathbf{W}_1^\intercal$$
$$\tilde{\mathbf{A}}_{1,i}, \tilde{\mathbf{A}}_{1,j} \text{ are independent}$$
$$\mathbf{K}_1 = \boldsymbol{\Phi}\boldsymbol{\Phi}^\intercal$$
$$\underbrace{\mathbf{A}_2}_{N\times p} = \underbrace{\mathbf{K}_2}_{N\times N}\underbrace{\mathbf{W}_2}_{N\times p} \quad \text{for } \mathbf{W}_2 \text{ kPCs s.t.}$$
$$\mathbf{W}_2^\intercal\mathbf{W}_2 = \mathbb{I}_p$$
$$\mathbf{W}_2^\intercal\mathbf{K}_2 = \boldsymbol{\Lambda}_2\mathbf{W}_2^\intercal$$
$$\tilde{\mathbf{A}}_{2,i}, \tilde{\mathbf{A}}_{2,j} \text{ are independent}$$
$$\mathbf{K}_2 = \boldsymbol{\Psi}\boldsymbol{\Psi}^\intercal$$

where $\boldsymbol{\Phi}$ and $\boldsymbol{\Psi}$ represent the non-linear maps applied to data sets $\mathbf{X}_{N\times d'}$ and $\mathbf{Y}_{N\times d}$, respectively. The matrices of $\mathbf{A}_{1_{N\times p'}}$ and $\mathbf{A}_{2_{N\times p}}$ are the matrices of the kPCA obtained from the kernel matrices eigen decomposition. Hence, this time in the second stage the CCA will be

obtained via matrices

$$\underbrace{\hat{I}\hat{I}^{\intercal}}_{p'\times p'} = (\mathbf{A}_1^{\intercal}\mathbf{A}_1)^{-1}\mathbf{A}_1^{\intercal}\mathbf{A}_2(\mathbf{A}_2^{\intercal}\mathbf{A}_2)^{-1}\mathbf{A}_2^{\intercal}\mathbf{A}_1$$

$$\underbrace{\hat{I}^{\intercal}\hat{I}}_{p\times p} = (\mathbf{A}_2^{\intercal}\mathbf{A}_2)^{-1}\mathbf{A}_2^{\intercal}\mathbf{A}_1(\mathbf{A}_1^{\intercal}\mathbf{A}_1)^{-1}\mathbf{A}_1^{\intercal}\mathbf{A}_2$$

Note that the columns of $\mathbf{A}_1$ and $\mathbf{A}_2$ are orthonormal. If now one considers the transformation introduced and rewrites the above then

$$\underbrace{\hat{I}\hat{I}^{\intercal}}_{p'\times p'} = ((\mathbf{K}_1\mathbf{W}_1)^{\intercal}(\mathbf{K}_1\mathbf{W}_1))^{-1}(\mathbf{K}_1\mathbf{W}_1)^{\intercal}(\mathbf{K}_2\mathbf{W}_2)((\mathbf{K}_2\mathbf{W}_2)^{\intercal}(\mathbf{K}_2\mathbf{W}_2))^{-1}(\mathbf{K}_2\mathbf{W}_2)^{\intercal}(\mathbf{K}_1\mathbf{W}_1)$$

$$\underbrace{\hat{I}^{\intercal}\hat{I}}_{p\times p} = ((\mathbf{K}_2\mathbf{W}_2)^{\intercal}(\mathbf{K}_2\mathbf{W}_2))^{-1}(\mathbf{K}_2\mathbf{W}_2)^{\intercal}(\mathbf{K}_1\mathbf{W}_1)((\mathbf{K}_1\mathbf{W}_1)^{\intercal}(\mathbf{K}_1\mathbf{W}_1))^{-1}(\mathbf{K}_1\mathbf{W}_1)^{\intercal}(\mathbf{K}_2\mathbf{W}_2)$$

Then, remark that $\mathbf{W}_1$ and $\mathbf{W}_2$ correspond to the kPCA projections for the two data set respectively, then

$$\underbrace{\hat{I}\hat{I}^{\intercal}}_{p'\times p'} = \Lambda_1^{-1}(\mathbf{K}_1\mathbf{W}_1)^{\intercal}(\mathbf{K}_2\mathbf{W}_2)\Lambda_2^{-1}(\mathbf{K}_2\mathbf{W}_2)^{\intercal}(\mathbf{K}_1\mathbf{W}_1)$$

$$\underbrace{\hat{I}^{\intercal}\hat{I}}_{p\times p} = \Lambda_2^{-1}(\mathbf{K}_2\mathbf{W}_2)^{\intercal}(\mathbf{K}_1\mathbf{W}_1)\Lambda_1^{-1}(\mathbf{K}_1\mathbf{W}_1)^{\intercal}(\mathbf{K}_2\mathbf{W}_2)$$

If $p = p'$, then

$$\underbrace{\hat{I}\hat{I}^{\intercal}}_{p'\times p'} = \Lambda_1^{-1}\Lambda_2^{-1}\Lambda_1(\mathbf{K}_2\mathbf{K}_2^{\intercal}) = \Lambda_2^{-1}(\mathbf{K}_2\mathbf{K}_2^{\intercal})$$

$$\underbrace{\hat{I}^{\intercal}\hat{I}}_{p\times p} = \Lambda_2^{-1}\Lambda_1^{-1}\Lambda_2(\mathbf{K}_1\mathbf{K}_1^{\intercal}) = \Lambda_1^{-1}(\mathbf{K}_1\mathbf{K}_1^{\intercal})$$

Hence, in the case of the kPCA-CCA, the coefficients $\mathbf{a}_i$ and $\mathbf{b}_i$ will be the eigenvectors associated to the highest canonical roots in the matrices $\Lambda_2^{-1}(\mathbf{K}_2\mathbf{K}_2^{\intercal})$ and $\Lambda_1^{-1}(\mathbf{K}_1\mathbf{K}_1^{\intercal})$ respectively. This concludes a detailed comparison between classical CCA and the novel framework proposed of reference method linear PCA-CCA and the non-linear version of kPCA-CCA. Clearly indicating the relationships between each method and the choice of projection basis and the kernel's influence on the CCA outputs.

5. Data and experiments

In our experimental studies, we focused on the U.S. state of California and utilised three distinct data sets. We engaged in extensive data sourcing to initiate the process, collecting relevant variables in these data sets. A crucial aspect of this work involved the engineering of unique features that require a high level of proficiency in advanced data processing, cleaning, and wrangling techniques.[3]

The veracity of the data sets utilised is outlined below:

1. **Pollution Air Quality Data**: ($\mathbf{X}^1_{N_1\times D_1}$) Sourced from the U.S. Environmental Protection Agency website (https://www.epa.gov/[4]), this data set provides a comprehensive view of environmental pollutants in multiple parameters and is sourced from the leading data agency for this data in the U.S.A.. The D_1 observed signals corresponding to carbon dioxide (Co2), nitrous oxide (No2), air quality (AQI), and particular matter 2.5 (PM2.5) observed daily between 2010–2020. The monitoring stations considered are selected within a maximum distance radius of 50 km from the main cities in Table 2. Fig. 5 shows four panels where purple reflects the selected cities and green shows the pollution monitoring station sites selected. Note that the top panels refer to No2 and Co2 (left and right, respectively), while the bottom panels refer to PM2.5 and AQI (left and right, respectively). Fig. 6 presents four barplots, showing how many stations are within each county. The top panels refer to No2 (left) and Co2 (right), while the bottom panels refer to PM2.5 (left) and AQI (right), respectively. Fig. 3 presents the feature extraction procedure for the S counties in the study, a daily time series per county is obtained by averaging the available monitoring station data for each pollution variable spatially each day, after removing monitoring stations with missing data. Fig. 7 demonstrates spatial heatmaps of the resulting pollution features averaged across 10 years by county. In the top panels, there are averaged No2 and Co2 (left and right respectively), while the bottom panels show averaged PM2.5 and AQI (left and right respectively) and Fig. 8 shows the boxplots of quarterly averages by county.
2. **Climate Data**:($\mathbf{X}^2_{N_2\times D_2}$) Sourced from the Global Surface Summary of the Day (GSOD) data set from the National Oceanic and Atmospheric Administration (NOAA) (https://www.ncei.noaa.gov/access/metadata/landing-page/bin/iso?id=gov.noaa.ncdc:C00516).[5] This data set offers a daily overview of global climatic conditions and is sourced from the leading data agency for this data in the U.S.A.. The D_2 observed signals corresponding to the mean temperature (in Fahrenheit) denoted MT, the maximum and minimum temperatures (in Fahrenheit) denoted H_t and L_t respectively and the total amount of precipitation (in inches), all variables are observed daily between 2010–2020. The monitoring stations considered are selected within a maximum distance radius of 50 km from the main cities in Table 2. Fig. 9 shows in the left panel a map of the selected weather stations in the State of California and in purple the selected counties are indicated and in green the weather stations around the counties are indicated and in the right panel the number of stations per county is displayed. In designing features for the climate variables, for every station in each county, first a cubic spline is used to interpolate any missing data. Then four spatial summary statistics time series are constructed per county. The first a bivariate time series of daily average high and low temperatures, each spatially averaged across all sensors in a given county. This produces a total of S bivariate daily average high and daily average low temperature time series between 2010–2020, i.e. one bivariate daily time series per county associate with selected cities in Table 2. The second set of feature time series extracted for the temperature data again uses the daily high and low temperature records, but transforms them into a volatility estimator based on Parkinson range based volatility (see [55]) which captures variation in temperature over time per sensor location. The volatility on day t for county s and monitoring sensor location l is given by

$$\sigma_{t,s,l} = \sqrt{\frac{1}{4n\times\log 2}\sum_{t=1}^{n}\left(\log\frac{H_t(s)}{L_t(s)}\right)^2} \tag{6}$$

Once the Parkinson volatility is obtained for every location, an average spatial volatility is calculated per day for each county. For the precipitation variable, a daily time series is obtained per county based on the total precipitation from a spatial aggregation of rainfall recorded within each station in a given county each day. Fig. 4 summarises the data preprocessing and feature

[3] The result of this extensive operation was the creation of three accessible data sets, available at https://github.com/mcampi111, which are invaluable assets for future research endeavours and reproducibility of results contained.

[4] Air Quality Data AQI https://aqs.epa.gov/aqsweb/airdata/download_files.html#Meta.

[5] Data is derived from the integrated surface hourly (ISH) data set. More information about such a dataset is provided at https://www.ncei.noaa.gov/access/metadata/landing-page/bin/iso?id=gov.noaa.ncdc:C00516.

extraction framework for the climate data. A summary of the features extracted is provided in Fig. 11 which demonstrates boxplots of the averaged features quarterly. In Fig. 10 spatial heatmaps of each climate feature averaged over 10 years for the mean temperature (left), averaged total precipitation (middle) and averaged volatility (right).

3. **Green Investment in Sustainable Development Goals via Municipal Green Bonds Data**:($\mathbf{X}^3_{N_3\times D_3}$) This data set was collected through a Bloomberg Terminal, providing comprehensive information on municipal green bonds issued within the state of California in the United States. This data was sourced from the leading commercial data service for financial data globally.

As the green bond market evolves and expands, it is well-positioned to provide the necessary funding for green projects, stimulating participation from various stakeholders in the transition towards a more sustainable economy. Therefore, the right selection of bonds and their attributes can provide insight into the current utilisation and potential for environmental impact that the overall green bond market may have in funding green initiatives that can reduce pollution and influence climate change in the longer term. Note that in order to solely focus on the highest quality issuers and credit ratings as well as the least potential for green washing we have selected the municipal green bonds, those issued by Californian state governmental authorities or public sector entities in California. As noted previously, California has the highest issuance of green bonds from any state in the USA and so is the ideal candidate for this case study. In this financial data set, N_3 corresponds to the number of green bonds issued across each county of California. The considered time span for issuance from 2015 to 2020, since the Green bond market is a nascent market that is growing exponentially during this time span.

The first challenge is to identify and then screen for the appropriate selection of financial municipal green bonds. To identify eligible municipal green bonds we used Bloomberg's search function 'SRCH' to screen for green bonds. This market screening function allows users to create customised lists of loans, government and corporate bonds, structured notes, municipal bonds, and preferred securities from the Bloomberg database. The bond selection criteria for Municipal green bonds (as of October 19, 2020) were carefully chosen to ensure a comprehensive dataset. Asset classes had additional options such as including private securities (all asset classes), consolidating duplicate bonds (REGS, 144 A, and STRIPs), including non-Bloomberg-verified bonds, and including strips (loans). The screening criteria utilised are presented in Tables 3, 4, and 5. A total of 1,425 municipal green bonds were issued between January 1st, 2015, and October 15th, 2020, within the US. Notably, we further filtered the 208 bonds related to California. We excluded bonds with adjustable and floating coupon rates to prevent estimation distortions. Each table presents selected bonds and additional statistical information from Bloomberg, including offer type (Negotiated or Competitive), underwriters, yield at issue, credit ratings and outstanding amount. Table 5 displays the selection criteria for Municipal green bonds (as of October 19, 2020) without optionality, resulting in 3,436 matches. This was then reduced to 167 Californian Greend Bonds.

Having identified the relevant green bond instruments in California, we then extracted the attributes recorded for these bonds. We applied filters using Bloomberg search fields, such as asset class, security status, environmental, social, and governance (ESG) green instrument indicator, issue date, maturity type, outstanding amount, and Bloomberg composite rating. The ESG criteria specified that the net proceeds of the fixed-income instrument should be applied towards green projects or activities promoting climate change mitigation or adaptation or other environmental sustainability purposes. While this ESG criterion applies to various types of bonds, including corporate bonds, preferred securities, and loans, the focus of the selection was on municipal bonds. For municipal bonds, the 'Y' (Yes) designation is returned if the bond has been classified as a green bond in either the municipal purpose (DS066, MUNIPURPOSE2) or the Municipal Purpose 3 (DS076, MUNIPURPOSE3) categories. It is important to note that the 'composite rating' criterion was not applied during the bond selection process. The collected financial variables offer valuable information about issued green bonds, aiming to describe each bond with relevant attributes that characterise it individually and convey information about the impact of its disbursements. In particular for each municipal green bond some of the key attributes collected included: issuance size, maturity, amount issued, yield at issue, coupon, spread, credit risk amongst other variables, see Table 6 which shows information about the collected variables, i.e. the name, a brief description and the data type.

Fig. 12 shows the number of issued green bonds per county within California. Alameda, San Francisco, Santa Clara, Santa Cruz and San Diego have significantly high numbers. Fig. 13 plots the feature embeddings for the green bonds after hot encoding is applied to the data from bonds in each county considered.

It should be noted that since the collected variables describing the green bonds were a combination of numerical, categorical and dates, cleaning procedures and hot encoding were employed to treat such differences when developing a feature representation. After cleaning and retention of the bond data that was complete, a total of 167 green bonds were retained across the counties of interest. Furthermore, since municipal green bonds tend to focus on expenditure of their use of proceeds within locations associated with the municipal issuer, it is relevant to group these instruments and their financial attributes according to the spatial location of the issuer, using this as a proxy for the location of the disbursement of funds to green projects. As such, the green bonds issued were categorised according to the geographical area associated with the county of issuance. We plot the location of the issuer for each retained green bond in Fig. 12.

With regard to encoding the non-numeric variables into a numeric feature space, several procedures could be followed for hot encoding. These include classical and contrast encoders, such as ordinal, one-hot, binary, hashing, Helmert, backward difference, polynomial, etc. Alternatively, one could consider Bayesian encoders such as target, leave-one-out, weight of evidence, James Stein method, M estimator, etc. The reader might refer to Cerda et al. [56] for a review of different hot encoding methodologies. In this work, for the categorical attributes, we use the most used in practise, corresponding to the one-hot encoding, which creates a new column for each unique value of the categorical variable. If, for example, a categorical variable has categories $\{\text{red}, \text{blue}, \text{yellow}\}$, the one-hot encoding will produce a three-dimensional feature vector defined as $\{[1,0,0],[0,1,0],[0,0,1]\}$. In the resulting vector space, each category is orthogonal and equidistant to the others. This property agrees with classical intuitions about nominal categorical variables in statistics. We perform a different solution for this encoding in the case of date variables. In practice, we take the minimum date across the entire data set and then count the number of days from that minimum date to the rest of the data. In such a way, the encoding is representative of the given data set and, thus, data-driven.

Each dataset is subject to a carefully designed data cleaning procedure and feature extraction process before being input to the methods

Table 2
Table providing the major cities in the US state of California with a population greater than 250,000 inhabitants.

Major cities California for monitor selection			
City	Population	Latitude	Longitude
Anaheim	334,909	33.84	−117.87
Bakersfield	301,775	35.36	−119.00
Fresno	472,517	36.78	−119.79
Long Beach	486,571	33.79	−118.16
Los Angeles	3,911,500	34.11	−118.41
Oakland	393,632	37.77	−122.22
Riverside	306,351	33.94	−117.40
Sacramento	480,392	38.57	−121.47
San Diego	1,299,352	32.81	−117.14
San Francisco	723,724	37.77	−122.45
San Jose	897,883	37.30	−121.85
Santa Ana	344,086	33.74	−117.88
Stockton	299,188	37.97	−121.31

explored for PCA-CCA and kPCA-CCA. A summary of the key aspects of data pre-processing and extraction of relevant features via the PCA and kPCA methods per data set is outlined in Fig. 2.

5.1. Data preparation and feature constructions

The following tables demonstrate the screening criteria and results obtained from Bloomberg when extracting municipal green bonds.

5.2. Pollution air quality spatial-temporal data features empirical analysis

Empirical spatial–temporal summaries of the Air Quality pollution data collected.

5.3. Climate spatial-temporal data features empirical analysis

5.4. Green bond spatial-temporal data features empirical analysis

An outline of the green bond instrument attributes collected that were transformed into features in the analysis.

5.5. Training of kernel hyperparameters in kPCA via pre-image methods

In order to compute the kPCA for each data set, one must first estimate optimal kernel hyper parameters, i.e., the γ length-scale parameter of the radial basis kernel function used in this analysis for real values attributes, generically given by

$$k^{(RBF)}\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right) = \sigma^2 \exp\left(\frac{\left(\boldsymbol{x}_i - \boldsymbol{x}_j\right)^2}{2\gamma^2}\right). \tag{7}$$

Note, throughout we set $\sigma^2 = 1$ as we pre-whitened feature vectors prior to kernel mapping. For the categorical and ordinal data attributes in the financial data set a Jaccard index kernel is used which for two discrete finite sample sets, the Jaccard index is the size of the intersection divided by the size of the union of the sample sets of features, such that for $\boldsymbol{x}_i \in A \subseteq \Omega$ and $\boldsymbol{x}_j \in B \subseteq \Omega$

$$k^{(J)}(A, B) = J(A, B) := \frac{|A \cap B|}{|A \cup B|}$$

with Jaccard index $J(A, B) = 1$ when both $|A| = |B| = \emptyset$. Note, for the Jaccard kernel there are no hyperparameters to be learnt.

Therefore, for the RBK kernel, the γ length-scale hyperparameter for each data set is estimated via the method of pre-images combined with a hyperparameter learning algorithm. This involved considering a grid of values for the hyperparmeters γ and then computing the distances between the obtained pre-image for every tested hyperparameter and every data set and minimising the Euclidean distance between the obtained pre-image and the original data (per data set, per county). Table 7 shows the final set of hyperparameters minimising the Euclidean distance for every data set. Note that the columns represent the three data sets and the considered counties' rows.

5.6. Extraction of kPCA eigen functions for financial mixed attribute data

Since the financial data set is comprised of mixed type feature attributes - real valued numerical, categorical, dates and ordinal data, it was important to carefully consider the encoding and to examine the best approach to kernel construction on this combined feature space. The approach we adopted was to focus one kernel on sub-space associated with one portion of encoded attributes and a second kernel on the remaining feature sub-space and to combine them in an additive manner. Conceptually, this is equivalent to taking a feature space with feature vector i and splitting it into two sub-spaces as follows $\boldsymbol{x}_i = \left[\boldsymbol{x}_{1,i}, \boldsymbol{x}_{2,i}\right] \in \mathbb{R}^d$ where $\boldsymbol{x}_{1,i} \in \mathbb{R}^{d'}$ and $\boldsymbol{x}_{2,i} \in \mathbb{R}^{d-d'}$ and then using this sub-space decomposition to form an additive combined kernel, one specifically for each sub-space, as follows:

$$k\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right) = k^{(RBF)}\left(\boldsymbol{x}_{1,i}, \boldsymbol{x}_{1,j}\right) + k^{J}\left(\boldsymbol{x}_{2,i}, \boldsymbol{x}_{2,j}\right).$$

Of course this can be done for any number of sub-spaces of sub-space projections, and in this work the choice has been made to opt for two specific sub-spaces associated with the two different categories of features in the financial data, so that a specific kernel can be considered for the categorical and ordinal encoded features, distinct from the real values quantities.

Hence, once the encoding of all the non-numerical variables is performed, we can construct a kernel matrix from which we can extract the eigen functions via kPCA. To perform this we utilise an additive kernel in which the first component kernel is applied to the feature sub-space associated with numerical and encoded date variables and the second additive component kernel is applied to the encoded categorical variables. For the first kernel, we use the radial basis function with euclidean distance; for the second kernel, we will use the Jaccard index kernel. This allows us to produce a combined kernel matrix. A summary of this process is presented in the heatmaps in Fig. 14, where the columns of the plot show matrix heatmaps corresponding to the first kernel component presented in column one, the second kernel component presented in the middle column and the combined kernel matrix presented in the third column on the right. Kernel matrices computed on the numerical and date attributes are in left panels and the kernels on the encoded categorical attributes are in the middle panels with combined additive kernel results for both in the right column. The rows of this plot of heat maps of kernel matrices correspond to different counties (spatial analysis) for the counties of Alameda, San Francisco, Los Angeles, Santa Cruz, San Diego. Afterwards, we summed the two Gram matrices and obtained the final matrix carrying the information of all the financial variables for one county.

5.7. Analysis of the PCA and kPCA indexes

The approach adopted to assessing the information content captured by each of the extracted indexes, we term PCs (eigen vectors) or kPCs (eigen functions), is to assess the information captured by the rank reduced representations of the kernel matrices versus the complete data covariance matrix as measured by the "Empirical Centered Kernel Alignment" (CKTA) [57]. To perform this CKTA measure, considers two Gram Matrices $\mathbf{K}_1$ and $\mathbf{K}_2$ and define the CKTA as

$$\hat{\rho}\left(\mathbf{K}_1, \mathbf{K}_2\right) = \frac{\langle \mathbf{K}_1^c, \mathbf{K}_2^c \rangle_{\mathcal{H}}}{\sqrt{\langle \mathbf{K}_1^c, \mathbf{K}_1^c \rangle_{\mathcal{H}} \langle \mathbf{K}_2^c, \mathbf{K}_2^c \rangle_{\mathcal{H}}}} \quad \in \quad [-1, 1] \tag{8}$$

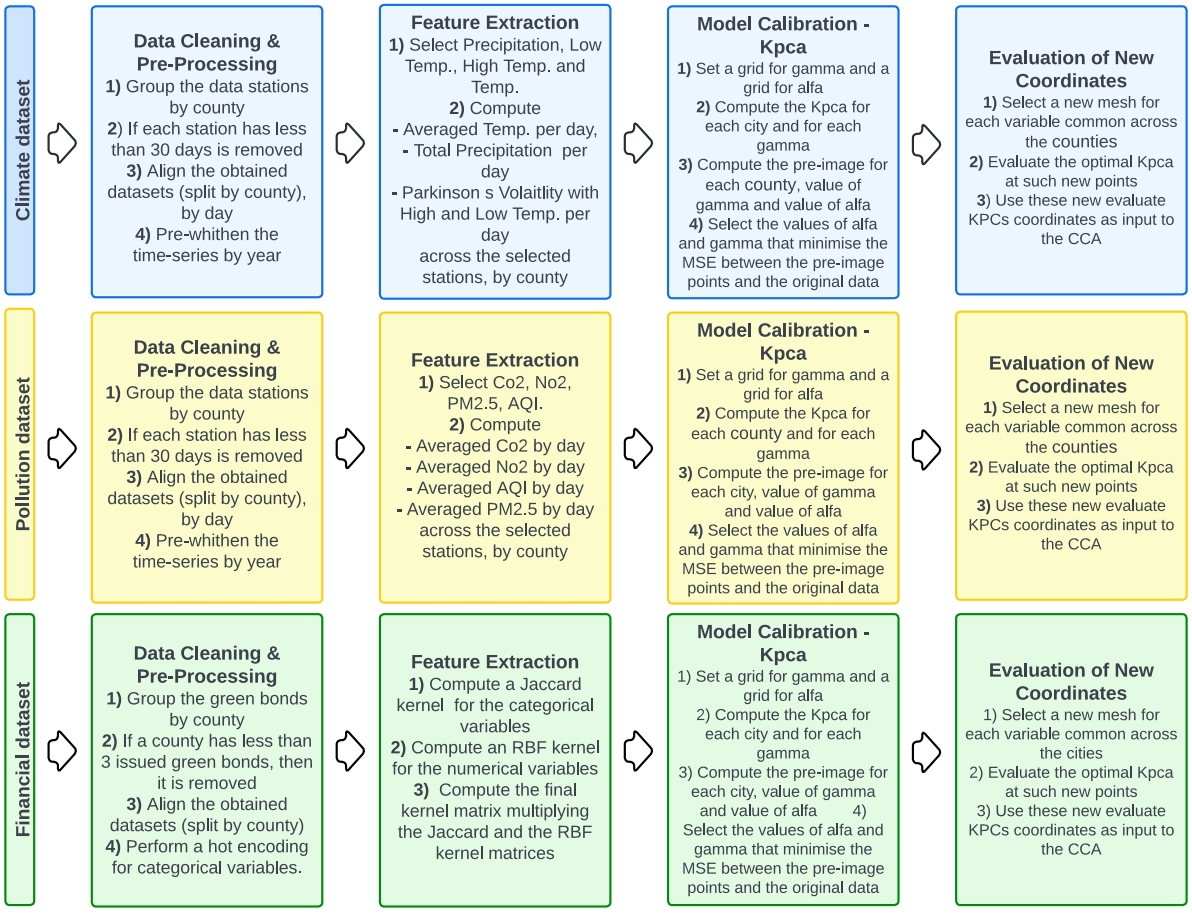

Fig. 2. The figure presents the steps performed on the different datasets. We have three data sets: climate, pollution, and financial ones. For each set, we performed data cleaning and pre-processing, a feature extraction procedure, model calibration for the kPCA with a grid search and the evaluation of a new set of coordinates by exploiting the out-of-sample problem. These steps have been introduced in Section 4.1 and will be further presented in this Section. The procedures are consistent across the data sets, but in the case of the financial data, in the feature extraction step, we consider two kernel functions: the radial basis function for the numerical variables and the Jaccard function for the categorical. More information about this will be given in the subsections below.

Table 3
Bond selection criteria for Municipal **conventional bonds** (as at 19-Oct-2020) without optionality.

Field	Boundaries	Selected criteria	Matches
Asset Classes	Include	Municiples	5,209,602
Security Status	Include	Active Municiples	947,379
Maturity Type	Exclude	Callable, putable, sinkable, make whole call, anticipated sinking fund	352,760
Issue date	In the range	01/01/2015–19/15/2020	273,824
Amount outstanding	≫	10 million	272,047
Composite rating	In between	AAA - BBB	XX

Table 4
Bond selection criteria for Municipal **green bonds** (as at 19-Oct-2020) with optionality.

Field	Boundaries	Selected criteria	Matches
Asset Classes	Include	Municiples	5,209,602
Security Status	Include	Active Municiples	947,379
Environmental, social & governance: green instrument indicator	Include		9,637
Issue date	In the range	01/01/2015–19/15/2020	8,839
Amount outstanding	≫	10 million	8,766
Composite rating[a]	In between	AAA - BBB	XX

[a] Criteria 'composite rating' is not applied.

where

$$\mathbf{K}^c = \mathbf{K} - \frac{1}{N}\mathbf{1}\mathbf{1}^\intercal\mathbf{K} - \frac{1}{N}\mathbf{K}\mathbf{1}\mathbf{1}^\intercal + \frac{1}{N^2}\left(\mathbf{1}^\intercal\mathbf{K}\mathbf{1}\right)\mathbf{1}\mathbf{1}^\intercal$$

corresponds to the centered kernel Gram Matrix, and $\mathbf{1}$ is the vector of ones with the appropriate dimension concerning the Gram Matrix $\mathbf{K}$.

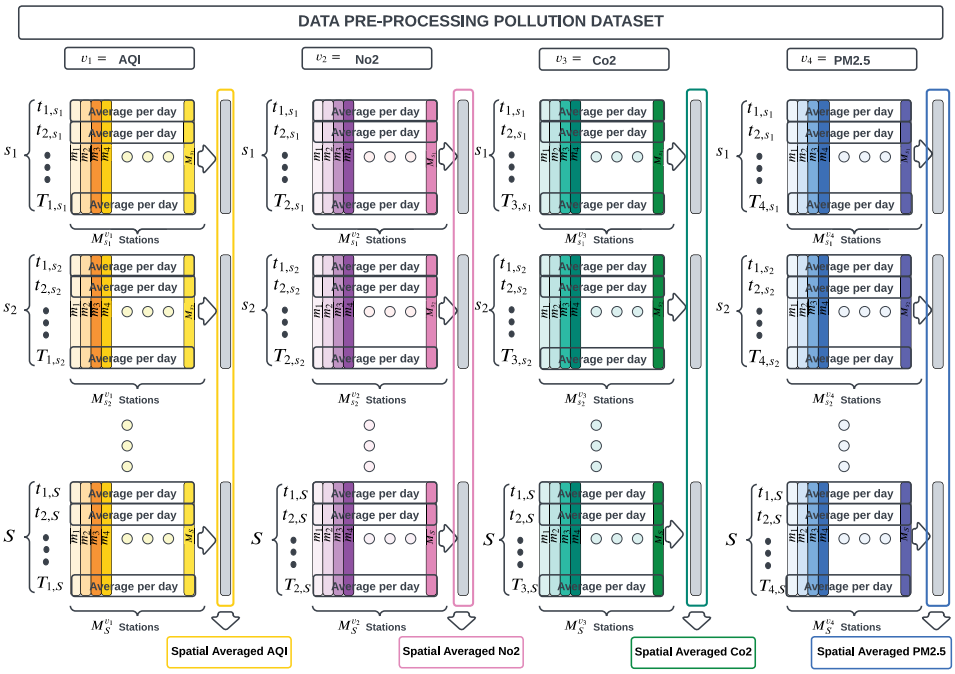

Fig. 3. Data cleaning and pre-processing of the pollution dataset. For each of the four variables and the S locations, we perform the following steps: (1) align the data at each monitor/station by day; (2) if data are missing, fit a cubic spline; (3) compute a time series spatial average across monitors by location. Full details of this procedure are provided in the initial part of this Section, under the pollution Air Quality Data description.

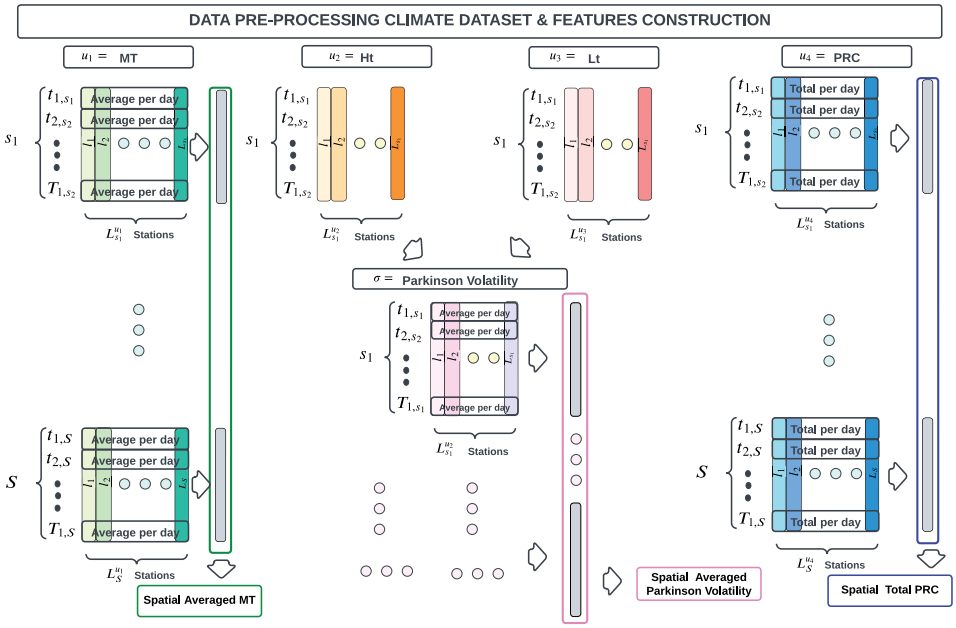

Fig. 4. Data cleaning and pre-processing of the climate dataset. For each of the four variables and the S locations, we perform the following steps: (1) align the data at each monitor/station by day; (2) if data are missing, fit a cubic spline; (3) compute a time series spatial average for different features. Full details of this procedure are provided in the initial part of this Section, under the climate Data description.

Furthermore, the operator $\langle \cdot, \cdot \rangle$ represents the matrix Frobenius inner product. Such an operator is computed on two real matrices $\mathbf{A} \in \mathbb{R}^{n \times m}$ and $\mathbf{B} \in \mathbb{R}^{n \times m}$ as follows

$$\langle \mathbf{A}, \mathbf{B} \rangle = A \cdot B = \sum_{i=1}^{n} \sum_{j=1}^{m} \mathbf{A}_{ij} \mathbf{B}_{ij} = \mathrm{tr}\left(\mathbf{A}^{\intercal} \mathbf{B}\right)$$

In the case of the PCA reduced rank approximations, the CKTA is utilised to compute the distance between the empirical covariance matrix and the kernel Matrices computed with one basis (PC1), with two bases (PC1 and PC2), and with three bases (PC1, PC2 and PC3). In the case of the kPCA reduced rank approximations, the CKTA is utilised to compute the distance between the empirical linear kernel data covariance matrix and the rank reduced kernel Matrices computed

Table 5
Bond selection criteria for Municipal **green bonds** (as at 19-Oct-2020) without optionality.

Field	Boundaries	Selected criteria	Matches
Asset Classes	Include	Municiples	5,209,602
Security Status	Include	Active Municiples	947,379
Environmental, social & governance: green instrument indicator	Include		9,637
Maturity Type	Exclude	Callable, putable, sinkable, make whole call, anticipated sinking fund	3,720
Issue date	In the range	01/01/2015–19/15/2020	3,474
Amount outstanding	≫	10 million	3,436
Composite rating	In between	AAA - BBB	XX

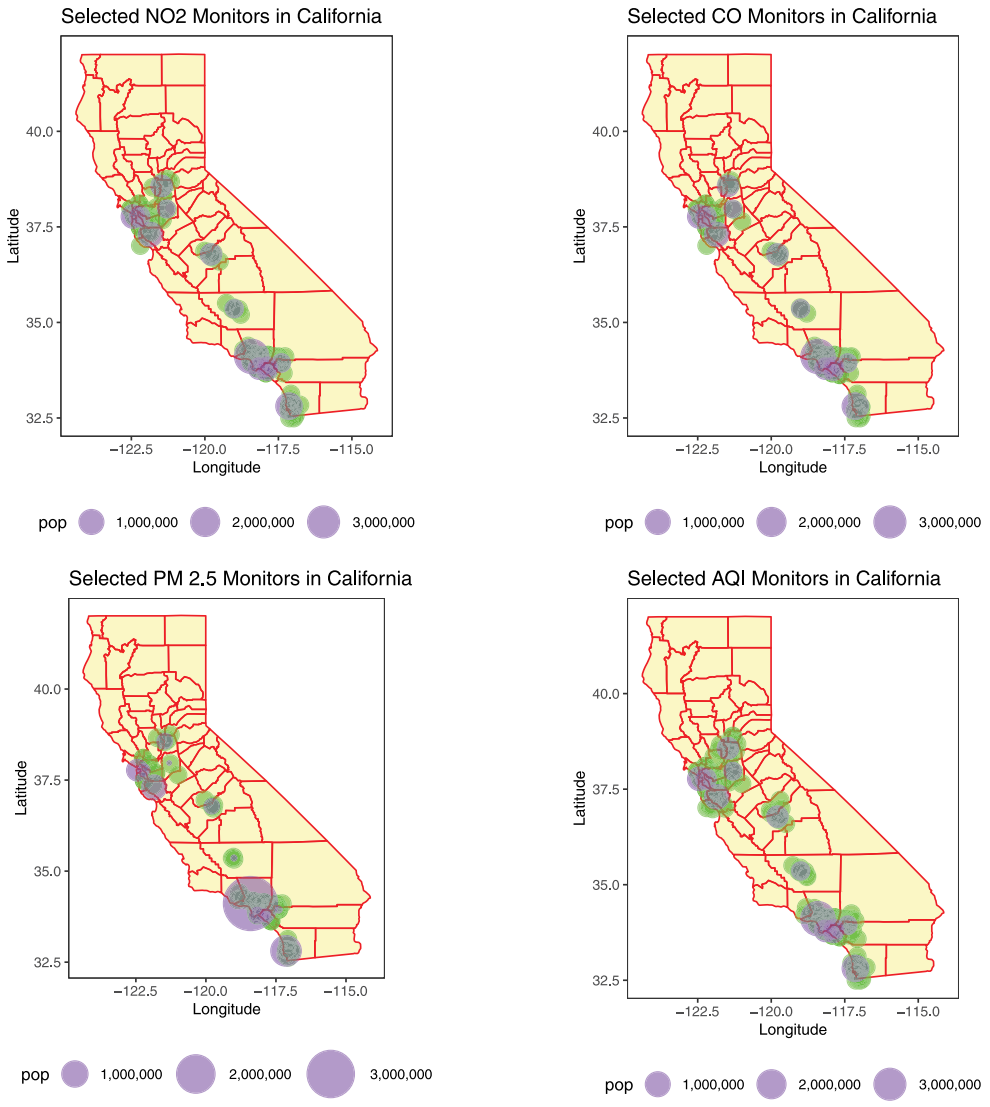

Fig. 5. Selected pollution monitors for collecting No2, Co2, PM2.5 and AQI. In purple major cities of California given in Table 2. In green the selected monitors which fall within a radius of 50 km around such cities. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

with one basis (kPC1), with two bases (kPC1 and kPC2), and with three bases (kPC1, kPC2 and kPC3). The results of this analysis allow us to order the rank reduced approximations, either linear PCA methods or non-linear kPCA methods according to their approximation accuracy as measured by the CKTA. This analysis is performed at a county level for every considered data set. In this way, we can assess which extracted basis functions best capture the structural variability of the proposed data features. Furthermore, by contrasting results for both PCs and kPCs, we can evaluate whether this variability is best represented by linear (PCA) or non-linear (kPCA) methods.

The CKTA results are summarised in Tables 8 and 9 for pollution and climate data sets, respectively. Optimal representations are highlighted based on the achieved CKTA scores that were superior to 70% as, in practice, a 70% level of alignment represents high variability captured. It is clear from this analysis that there is a definite advantage to using the extracted representations based on the kPCA eigen functions when

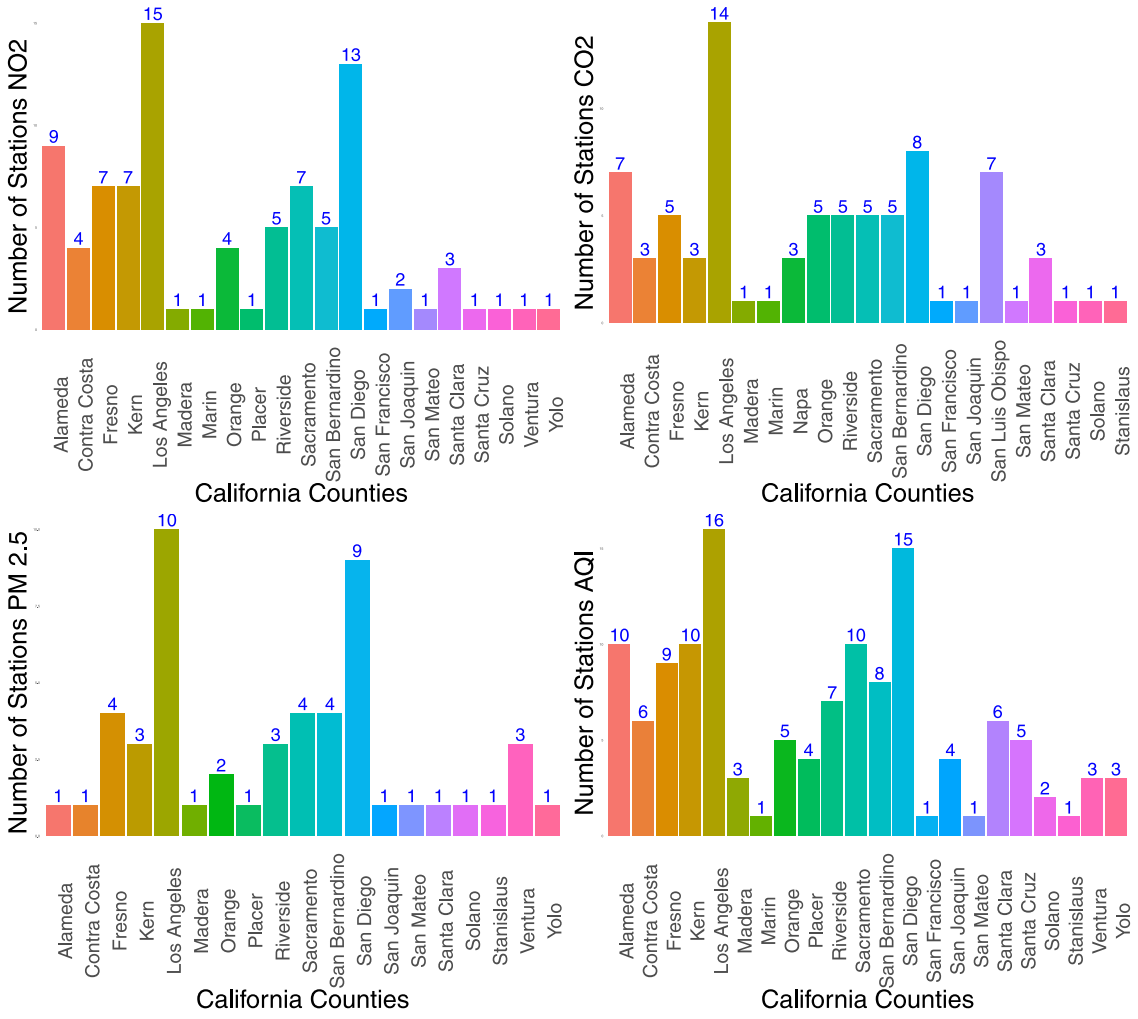

Fig. 6. Barplots showing the number of selected stations within each county of interest. Note that the top panels refer to No2 and Co2 (left and right, respectively), while the bottom panels refer to PM2.5 and AQI (left and right, respectively). The *x*-axis shows the different California Counties (ordered alphabetically from left to right), and the *y*-axis represents the number of stations considered for that variable. In the final experiments, only a subset of all these counties will be used. Further explanation is given in the following Subsections.

contrast with the PCA eigen vector bases. This is most pronounced on the Pollution Air Quality Data sets analysis, where PCA representations were significantly inferior to the kPCA representations across all spatial regions. In the case of the Climate data the effect was significantly less pronounced difference, though it is still the case that the kPCA was optimal across all counties.

The results for the financial data are presented in Table 10. In this case, the PCs and the kPCs are applied to the given data after performing hot encoding and ad hoc transformations for the kPCS described above but without engineering new specific financial features. The results show that PC alignments with the empirical covariance matrices do not achieve 70% in any of the counties, suggesting that the underlying data carries highly non-linear and non-stationary variability.

This indicates that indeed the modes of variation in each data set are best captured by the non-linear kernel basis representations. Validating the need to proceed to construct the kPCA-CCA method between pairs of data sets:

- Pollution Air Quality Data ($\mathbf{X}^1_{N_1\times D_1}$) and Green Investment in Sustainable Development Goals via Municipal Green Bonds Data ($\mathbf{X}^3_{N_3\times D_3}$); and
- Climate Data:($\mathbf{X}^2_{N_2\times D_2}$) and Green Investment in Sustainable Development Goals via Municipal Green Bonds Data:($\mathbf{X}^3_{N_3\times D_3}$).

Note, that whilst the PCA basis representations are sub-optimal in these findings compared to the kPCA representations, we will still develop solutions based also on PCA-CCA which will act as a baseline reference to compare performance to when assessing the preferred method of kPCA-CCA. A toy example describing the procedure of how to compute these CKTAs is provided in the Supplementary Information.

5.8. PCA-CCA and kPCA-CCA inter data analysis

This section focuses on the analysis of the PCA-CCA and kPCA-CCA results. Statistical tests, discussed in Section 3.3, that allow one to study the statistical evidence for the relationships extracted by the CCA component of the framework will be performed in order to assess the results of the canonical correlation for PCA-CCA and kPCA-CCA method results when applied to the pairs of data (Pollution Air Quality, Green Bonds Financial) and (Climate, Greend Bonds Financial) data. Subsequently, we will look at the canonical correlation by considering the structure coefficients or canonical loadings.

Furthermore, in the case of the kPCA features, in order to assess the correlation between green bonds' financial variables and pollution or climate and provide support to more efficient decision-making processes in how the use of proceeds is employed and, possibly, how this should be used in the future, we require a description of the contribution of the original financial variables to understand which one drives

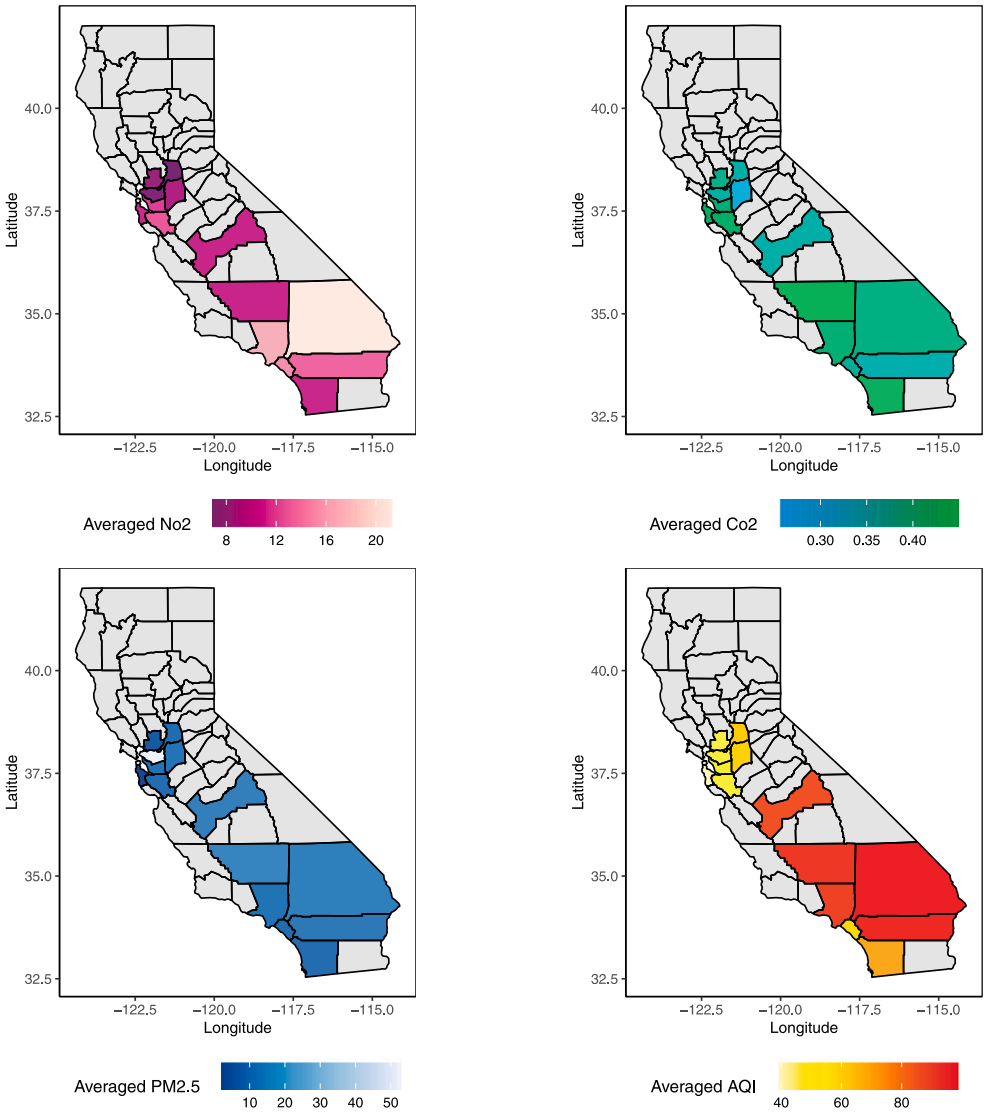

Fig. 7. Heatmaps of the engineered pollution features averaged across 10 years by county. In the top panels, there are averaged No2 and Co2 (left and right respectively), while the bottom panels show averaged PM2.5 and AQI (left and right respectively). We can observe the spatial distribution of the averaged features by monitors further averaged by 10 years.

the obtained results. This is achieved by considering a reconstruction error for the numerical financial variable, i.e. by taking the pre-image of the kPCs. We then computed the difference between the original and reconstructed data through the kPCs. For the categorical variables, instead, we employ the cKTA and observe, variable by variable, the alignment with the original categorical data. In other words, we select one county and one categorical variable, construct an empirical covariance matrix of that individual variable through the Jaccard distance, and, afterwards, compute the empirical covariance matrix of all categorical variables of that county (again with the Jaccard distance) and calculate the distance between the two matrices. As a result, we have an alignment per variable with that county's whole categorical data set and can interpret how much each variable contributes to the variation. All the analysis is done by considering the first two bases of each decomposition method, i.e. PC1 and PC2 and kPC1 and kPC2, since these two presented significant canonical correlations, while the third bases did not carry any. Further, note that the Supplementary Information presents one section with the analysis of the correlation matrices for the PCA-CCA and kPCA-CCA.

Results are provided in Tables 11(a) to 11(d) for the PCA-CCA, while in Tables 12(a) to 12(d) for the kPCA-CCA. Each table shows the main results with the canonical coefficients per canonical variates, the squared canonical correlation and the F-test for canonical correlation following Rao's approximation, for which we provide the F-statistic, the two degrees of freedom required for the computation of the test and the p-value. If one focuses on the first set of Tables, i.e. Tables 11(a) to 11(d), the top table refers to the CCA conducted with PC1 and the bottom tables with the CCA carried with PC2. Further, the left tables are for the financial/pollution CCAs, while the right tables are for the financial/climate CCAs.

5.8.1. Linear benchmark PCA-CCA inter data strength of spatial-temporal associations analysis in PCA factor space

The results in Tables 11(a) to 11(d) of application of the PCA-CCA analysis demonstrated that no canonical correlation is strong enough to be statistically significant. This was found to be true for any selected PCs according to the F-test. This demonstrates that the linear method for the feature extraction via PCA failed to capture significantly any evidence of associations between the changes in the green bond financial data variables and the pollution air quality both spatially and temporally. Likewise this was also found to be the case for the result seeking associations between the changes in the green bond financial

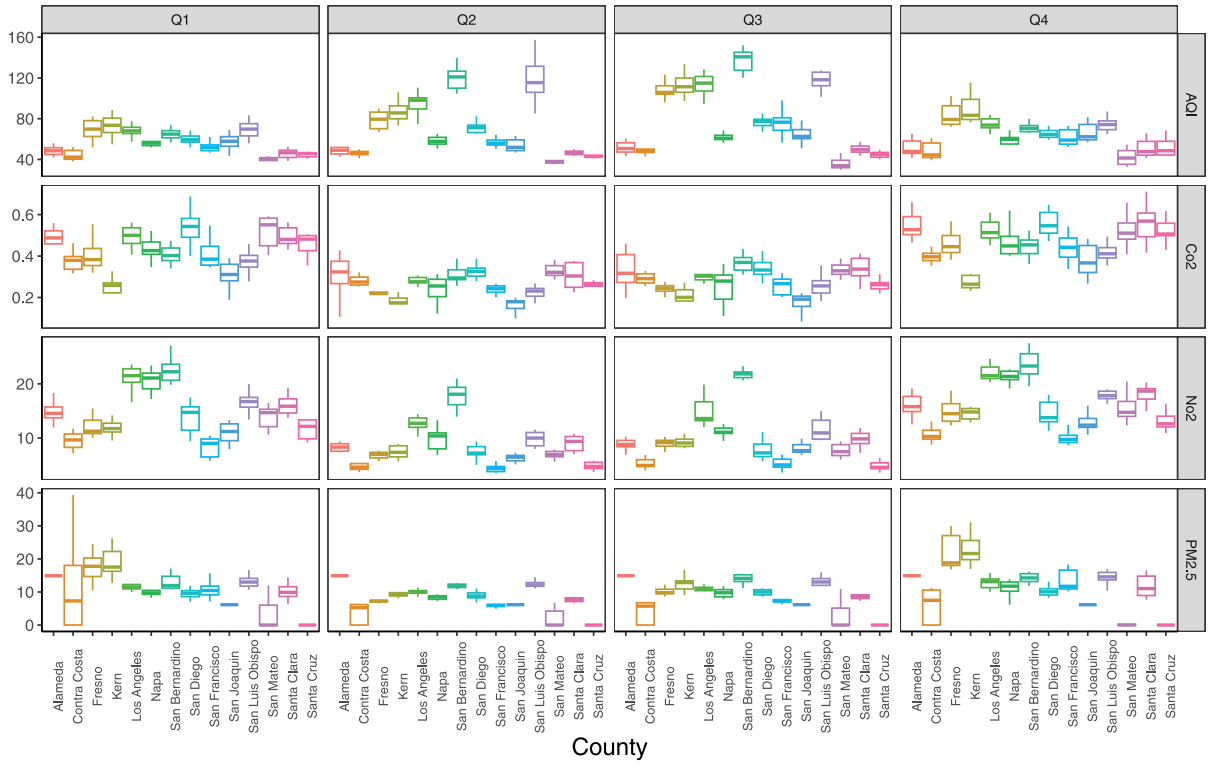

Fig. 8. Boxplots of the pollution-engineered features, further averaged by yearly quarters and county. Hence, each boxplot is representative of 10 points, where that variable has been averaged across the quarter and the county.

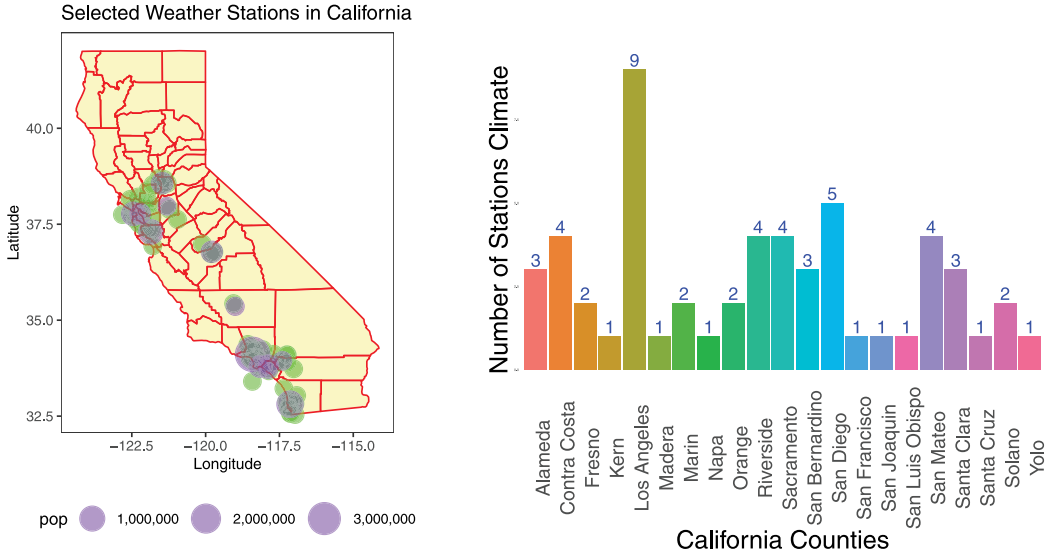

Fig. 9. Left panel: map of the selected weather stations in the State of California. In purple major cities of California given in Table 2. In green the selected monitors which fall within a radius of 50 km around such cities. Right panel: number of stations per counties. The x-axis shows the different California Counties (ordered alphabetically from left to right), and the y-axis represents the number of weather stations. In the final experiments, only a subset of all these counties will be used. Further explanation is given in the following Subsections. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

data variables and the climate data, spatially and temporally. It is conjectured that this lack of statistical evidence for such relationships arises from the fact that the variation in both the pollution air quality spatial–temporal process and the climate spatial temporal processes are best capture by non-linear features that can only be obtained via the kPCA method. Further evidence of this, as it pertains to assessment of the PCA-CCA method, is confirmed in Section 7 of the Supplementary Information.

5.8.2. Non-linear kPCA-CCA inter data strength of spatial-temporal associations analysis in kPCA factor space

Results for the non-linear kPCA-CCA methods application using kPCs instead of PCs are provided in Tables 12(a) to 12(d). Analogously

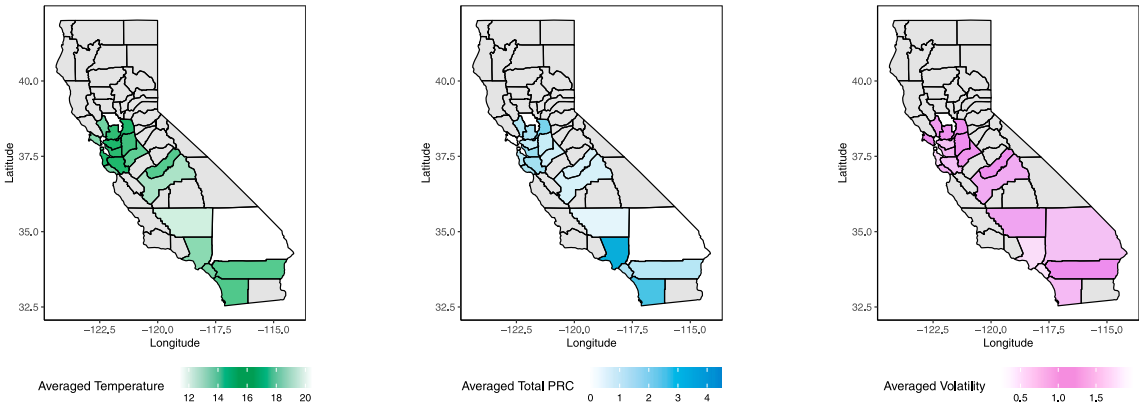

Fig. 10. Heatmaps of the engineered weather features averaged across ten years by county. From left to right, the panels show averaged temperature, averaged total precipitation and averaged volatility. We use the term averaged for each feature, meaning that the plotted value for every county is an average across the ten years.

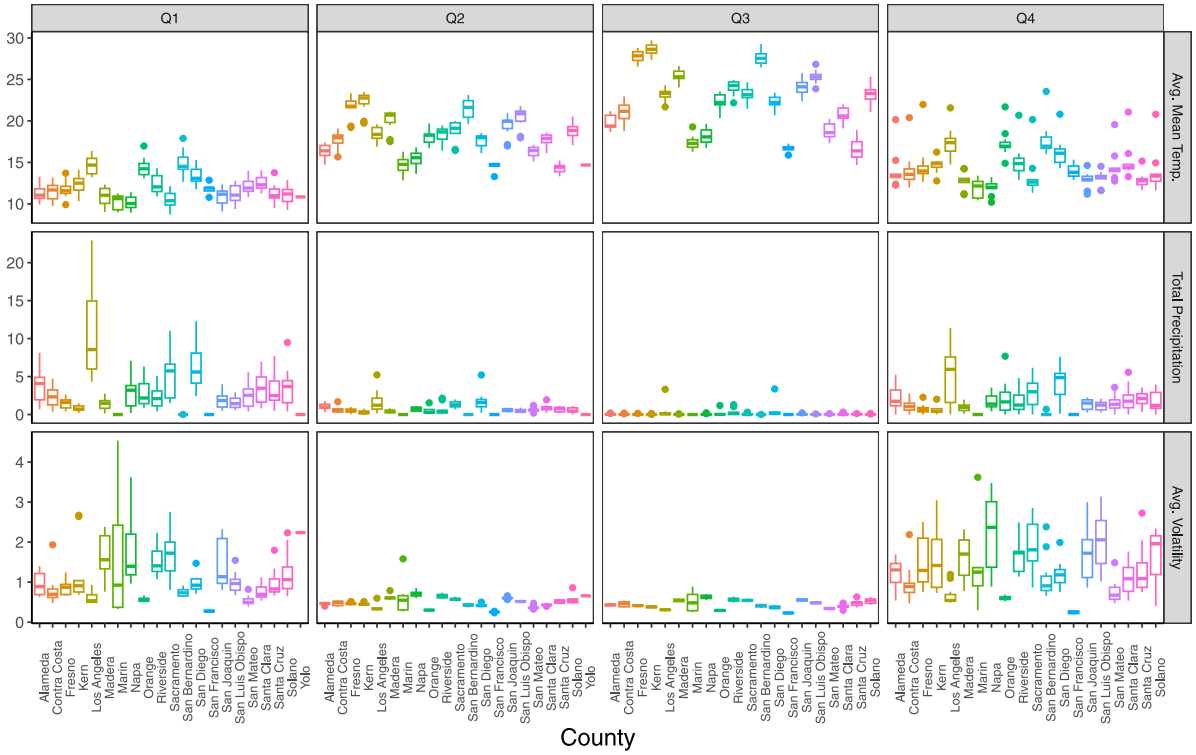

Fig. 11. Boxplots of the climate-engineered features, further averaged by yearly quarters and county. Hence, each boxplot is representative of 10 points, where that variable has been averaged across the quarter and the county.

to the PCA-CCA method, the results are presented for both kPC1 and kPC2, respectively. In the results for kPC1, the first two canonical functions display a canonical correlation of 1.000 and 0.790, which are high levels of correlation (note that we will consider across the entire set of results correlations that are superior to 0.700). Furthermore, the correspondent squared canonical correlations, representing the shared variance in each canonical function by the individual canonical variates, are 0.999 and 0.724, respectively. This suggests that these two canonical functions detect most of the underlying cross-correlation between the kPC1 extracted by these two data sets. All the canonical functions are significant according to the F-test, with the only two exceptions for the eighth and ninth. However, the first two will be considered in the analysis since they carry the highest level of correlations.

If one now considers the result of kPC2 given in Table 12, it is apparent that in this case also the canonical correlations of the first two canonical functions are strong and correspond to 1.000 and 0.881, respectively, with squared canonical correlations of 0.999 and 0.776. This demonstrates that high levels of variance are explained by both the canonical variates of each canonical function. Again, all the canonical functions except for the ninth one are significant according to the F-test. However, only the first two will be retained and considered in the analysis.

In Tables 12(c) and 12(d), equivalent results of the kPCA-CCA method are presented for kPC1 and kPC2 of the climate and green bond financial data. Unlike the case for the pollution air quality data versus green bond financial data, in this pair of data (climate and green bond financial data), the only canonical correlation higher than 0.700 is one of the first canonical functions of kPC1 with a canonical correlation

Table 6
This table lists the financial characteristics collected for each green bond, including a description of the attribute and its type, which may be categorical, numerical, or date type.

Financial data		
Variable	Description	Categorical/Numerical/Date
CUSIP	Committee on Uniform Security Identification Procedures Security identification number for the U.S. and Canada	Categorical
County of Issuance	County where the green bond is issued (in California)	Categorical
Issuer Name	Name of the issuing entity of the security.	Categorical
Muni Maturity Size	Dollar amount of bonds issued under this maturity. For Zero Coupon Bonds, the dollar amount represents the initial principal value.	Numerical
Amount Out	The total or principal amount of the green bond that has been disbursed or provided to the borrower or issuer	Numerical
Coupon	Current interest rate of the security. For bonds with reset compounding structures, this will return the estimated annualised daily reset compounding structures, this will return the estimated annualised rate for coupon cash flow calculations for the corresponding settlement date	Numerical
Issue Date	Date the security is issued	Date
Dated Date	Date when interests start to accrue	Date
Maturity	Date, the principle of a security, is due and payable	Date
Bid OAS (option adjusted spread) Spread (bps)	Number of basis points the spot curve would have to shift for the present value of the cash flows to equal the security's price, using the bid price	Numerical
Mid-Macaulay Duration	Macaulay's Duration based on the mid-price of the security is returned	Numerical
Issue Price	Price of the security at issue	Numerical
Yield at Issue	Occurring on the coupon strip's maturity date. Therefore, the amount outstanding/issued is not populated. Municipals - Returns the amount of the given maturity.	Numerical
Spread at Issuance to Worst	Spread for tax-exempt bonds is calculated from AAA Callable. For taxable bonds, the spread is calculated from US Treasury Actives curve. Spread is calculated to the appropriate interpolated point on the curve.	Numerical
Muni Issue Type	Describes the security structure of the bonds and the security type	Categorical
Issuer Industry	The industry classification of the issuer of the security	Categorical
Muni Source	The source of funds that will be the primary source of debt service on the bonds	Categorical
Muni Offering Type	Specifies how a bond was sold in the market. Bond sale methods can be competitive or negotiated. Short-term deals are typically 18 months or less in maturity. Limited sales are to a specific set of investors, while private placements are sold directly to investors with certain restrictions. Remarketed bonds are resold after they have been tendered	Categorical
Muni Issue Type	Describes the security structure of the bonds and the security type	Categorical
Bloomberg Issuer 5-Year Credit Risk	Risk class assigned to the issuer based on the on the Bloomberg Issuer Default Risk model generated probability of default over the next five years	Categorical

of 0.815 and a squared canonical correlation of 0.712. All the others, for both kPC1 and kPC2, are below this target 0.7 threshold, hence suggesting a low level of correlations between these two modes of variations extracted on financial data and climate data. Furthermore, by focusing on the F-test results, less canonical functions appear to be significant compared to the pollution/green bond data analysis case. The result section will therefore focus on pollution more than the climate but still show the obtained results for both cases to support our findings further.

Furthermore, it is interesting to observe how the rate of the canonical correlation decreases across the variates at a much slower pace compared to the one of the squared canonical correlation, hence suggesting that a researcher should be carefully paying attention to both these indices since even if the correlation is maximised, and the F-test appears to be significant, if the variance shared between the different synthetic canonical variates is low, then the correspondent pair or canonical function will not carry enough information of the underlying data.

At this stage of the analysis, a common practice is to consider the redundancy index. The redundancy index provides an indicator summary of the overall explanatory power of the canonical functions. In practice, it is to determine how much of the variance is accounted for in one set of variables by the other set of variables. We provide a redundancy analysis in the Supplementary Information, for kPC1-CCA and kPC2-CCA for the case of pollution and financial data. The redundancy plots show that the total variance of the financial kPCs explained by the corresponding pollution kPCs is approximately 60%, while the total variance of the pollution kPCs explained by the corresponding financial kPCs is about 80%. High redundancy suggests a strong predictive ability, indicating that green bond financial kPCs

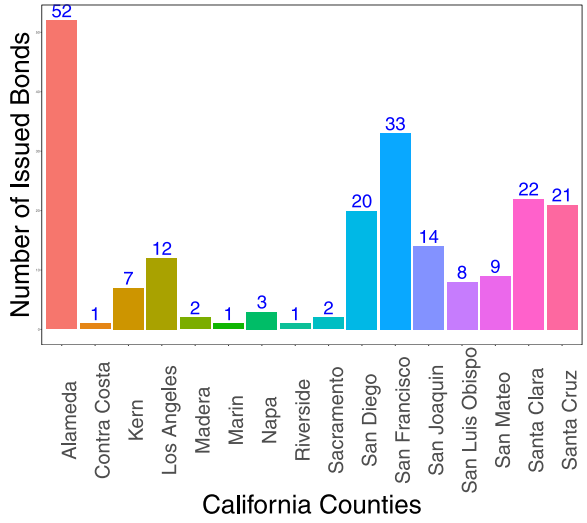

Fig. 12. Number of green bonds issued in California. Note that the total corresponds to 208. However, in the analysis, due to some data cleaning criteria the final number will be 167.

Table 7
Table describing the optimal γ parameters of the RBF kernel as given in Eq. (7). The procedure to identify these final hyperparameters is summarised Subsection 3.1.1 of the Supplementary Information. As explained, the kPCA is conducted at a county level, hence the first column presents the set of counties considered in California taken into account according to data availability of the three datasets and population number of the considered counties (note that this selection criterion information is explained in details in Section 5). The columns represents the three different data sets, i.e. financial data set, pollution data set and climate data set.

Optimal γ hyperparameter for the datasets			
County	Financial	Pollution	Climate
Alameda	0.5	0.5	5
Los Angeles	0.5	0.5	1
Napa	1.0	1.0	1
San Diego	0.5	0.5	5
San Francisco	0.5	0.5	5
San Joaquin	0.5	0.5	1
San Luis Obispo	0.5	0.5	5
Santa Clara	0.5	0.5	5
Santa Cruz	0.5	0.5	1

efficiently explain pollution kPCs. This suggests that increasing growth in green bond industry as measured by increased attributes such as issuance of greenbonds, increases issuance sizes that lead to greater funding for green initiatives is directly able to predict changes in pollution air quality. An equivalent analysis was performed for the case of financial and climate data, but no significant results were obtained.

In principle, the weak associations between the climate and green bond financial data sets, we believe is a result of the fact that the time-scale taken for green finance expenditures and funded projects to materially impact the climate in the regions in which significant funding is provided to green initiatives will be of a much longer time resolution to the time window studied in this work. Note, the time utilised in this project reflects the longest time possible as the green bond market is still in its infancy and so bonds have not been issued yet for multiple decades in this nascent market. As this market continues to grow and mature, we suspect that a re-assessment of this pair of data will yield stronger relationships from the kPCA-CCA methodology.

This analysis shows that whilst the emerging green bond market expenditure on green initiatives is having strong spatial–temporal associations with changes in pollution as a result of efforts to undertake green finance funded mitigation, this is detectable at shorter time-scales. However, the influence such green bond expenditures will ultimately have on climate variables will be a much longer process of assessment and will require multiple decades of green bond financing data to be definitive as to how effectively one can find associations between measurable changes in climate and green bond financing of projects to attempt to implement climate change mitigation strategies.

Next the results of the kPCA-CCA methods application is assessed in more spatial detail, which is achieved by focusing on the structured coefficients and the squared structure coefficients of the first and second canonical functions for the case of kPC1 and kPC2 for each pair of data sets analysed. These results are given in Tables 13(a) and 13(b), where the left tables refer to pollution air quality and green bond financial while the right tables refer to climate and green bond financial data, respectively. Further, the top tables refer to the results of kPC1 and the bottom tables to the results of kPC2. To further understand the output of these tables, we provide plots of the structured coefficients in Fig. 15 and Fig. 16, for pollution air quality and green bond financial; and climate and green bond financial, respectively. Each table shows the data set of interest, the kPC on which the CCA has been performed, the county for which we collected the structured coefficient of the first and second canonical functions and the squared structured coefficient of the first and second canonical functions.

It is known that a structured coefficient represents the equivalent interpretation of a loading in PCA and the bivariate correlation between an observed variable and a computed canonical variate. They range between -1 to 1 and provide information about which of the original variables, the kPCs, must define the canonical variate to maximise the correlation across these. Hence, how much the original quantities contribute or load the constructed canonical variate. Furthermore, the squared structure coefficient represents the proportion of variance an observed variable, hence a kPC, shares with the canonical variate generated from the CCA. The structured coefficient can be considered if this quantity is high enough. We will consider a structure coefficient higher than 0.700 or lower than -0.700 with a squared structure coefficient in an equivalent range. To analyse these tables, the analysis first considers the relationship between structure coefficients (and related squared structure coefficients) related to kPCs of the same data set and, after, across the different data sets. Consequently, a relationship between the different counties through the modes of variations given by the kPCs can be considered. The significant results and the ones discussed in the tables are highlighted in bold.

If one focuses on the top panel of Table 13, then the results for the CCA applied to the first kPC of financial and pollution data sets of the nine considered California counties are shown. Note that there are significant results only for the first canonical variate. A visual representation of these findings is also provided in a heliogram plot in Fig. 12 which shows the results per county. It is possible to observe how San Diego, San Francisco, Santa Cruz and Santa Clara have many issued green bonds compared to San Luis Obispo, so a different behaviour is expected. This can be found in kPC2.

We now move to the analysis of CCA on kPC2 for pollution and financial data sets. One observes that while the kPCA-CCA analysis using kPC1 appears to capture a multivariate relationship most strongly in spatial regions Alameda, Napa, San Joaquin and San Luis Obispo, in contrast, the application of kPCA-CCA using the second kPC instead presents a relationship also across other counties including San Diego and San Francisco, suggesting that these two modes of variations capture different underlying information. If one then looks at the green bond issuance data in these various counties this distinction is indicative of the difference in rate and size of issuance between these

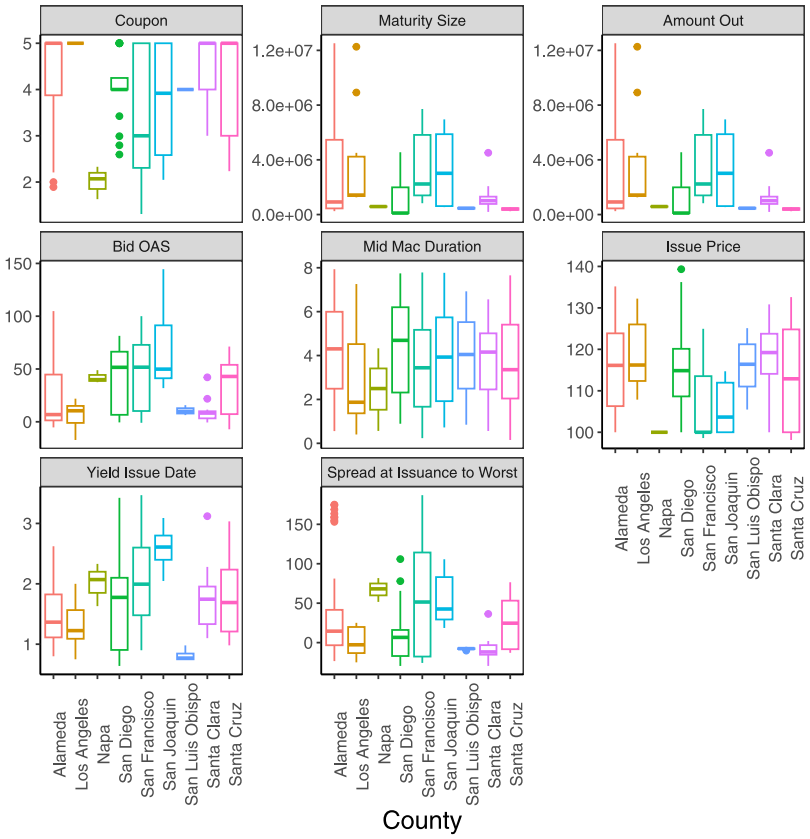

Fig. 13. Boxplots of the numerical variables used for the financial data set. The x-axis shows the considered Counties while the y-axis the range of the values for every variable.

Table 8
cKTA results for the pollution data set. Each row shows a considered county, while, in the columns, we have the different approximation matrices used for the cKTAs. The first column presents the cKTAs calculated using the covariance matrices of the engineered features for the pollution data and the rank-one approximation covariance matrices using PC1 as a column vector. The second column presents the cKTA calculated using the covariance matrices of the engineered features for the pollution data and the rank-two approximation covariance matrices using PC1 and PC2 as column vectors. Equivalent reasoning applies to the rest of the columns.

Results of centered kernel target alignment - pollution dataset

County	PC1	PC1 & PC2	PC1 & PC2 & PC3	kPC1	kPC1 & kPC2	kPC1 & kPC2 & kPC3
Alameda	0.160	0.476	0.536	0.276	0.479	**0.709**
Los Angeles	0.370	0.452	0.465	0.324	**0.740**	**0.808**
Napa	0.015	0.141	0.143	0.440	0.651	**0.803**
San Diego	0.280	0.530	0.551	0.262	**0.730**	**0.849**
San Francisco	0.656	0.624	0.695	0.568	0.553	0.511
San Joaquin	0.670	0.499	0.544	0.555	0.681	**0.719**
San Luis Obispo	0.072	0.411	0.459	0.660	**0.751**	**0.810**
Santa Clara	**0.766**	**0.799**	**0.806**	0.681	**0.714**	**0.886**
Santa Cruz	0.313	0.313	0.515	0.444	0.601	**0.705**

counties. Further results are provided in the Supplementary Information which show additional plots of structured coefficients for more canonical variates (the second one and the third one).

Table 13 shows results for the structured coefficients and squared structured coefficients of the first and second canonical variates of kPC1 and kPC2 for studying climate and green bond financial data. Results found here are consistent with the conjecture posed earlier that pertains to the inability to detect strong associations between green bond financing of climate mitigation strategies over the relatively short time frame studied, when compared to the time scale it will likely take for measurable associations between such green project expenditure and measurable climate mitigation to manifest. We simply include these results here for completeness of the analysis.

5.8.3. Interpretation of non-linear kPCA-CCA results in original data feature space

In this section, building on the results that identified strong associations in space and time between green bond financial data when projected into non-linear kPCA factor space vs pollution air quality data when also projected into non-linear kPCA factor space, it becomes interesting to re-assess these relationships identified in the original data

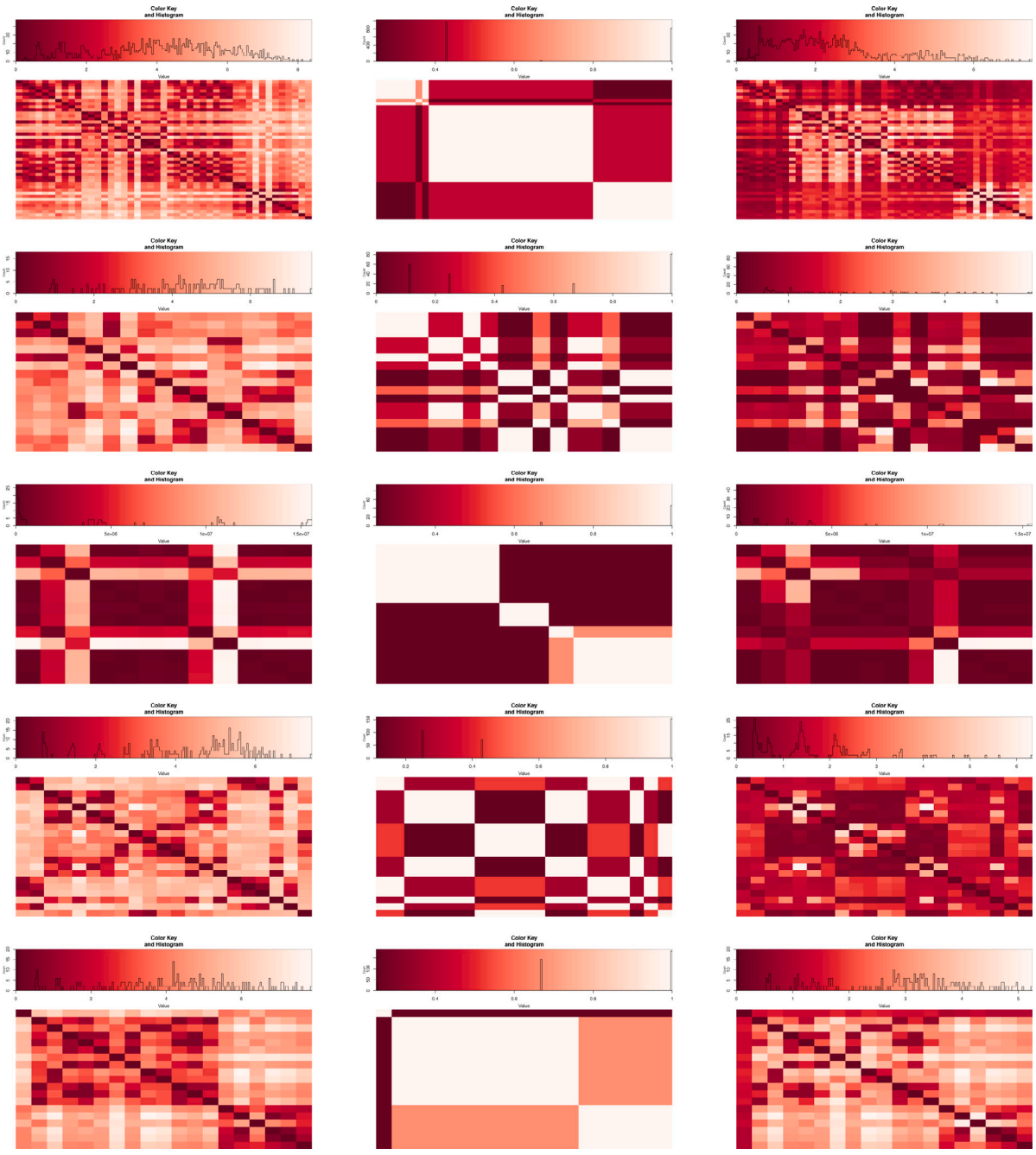

Fig. 14. Heatmaps of the Gram matrices computed with the Jaccard distance on the financial variables of the counties Alameda, San Francisco, Los Angeles, Santa Cruz, San Diego, from top to bottom. Note that within each individual panel, the top bar represents the colour key or legend. It represents the mapping between the colours used in the heatmap and the corresponding numeric values. The line within the colour key represents the range of values present in the heatmap, with the colours on the heatmap corresponding to different ranges of values.

feature space. This is instrumental in both direct interpretation as well as development of actionable decision making outcomes based on the findings.

In this work, what must be considered at this point is that we should evaluate the information captured by the kPCs extracted on the financial data to achieve such a goal. To do so, since the kPCs incorporate information of both numerical and categorical data, we decided to observe the cKTA of the kPCs with the original variables, one by one. In so doing, one may identify what the detected associations mean with regard to the original data features. This can be achieved by considering a reconstruction mean square error (MSE), where, by considering the pre-images of kPC1 and kPC2, we computed the euclidean distances of the reconstructed data and the original ones and utilised them in the cKTA measures evaluation. Figs. 17 and 18 display the cKTA measures by variable and by kPC, per county.

Fig. 17 shows the analysis for the categorical green bond financial variables. It demonstrates that strongest results were obtained for the counties of San Diego, Santa Clara, and Santa Cruz. In San Diego, the best represented variable corresponds to Muni Source (the issuer of the municipal green bond) for both kPCs, while, in Santa Clara the best categorical variable is Muni Issue Type (the structuring of the green bond). For Santa Cruz, the variables that were most influential were Muni Source, Muni Offering Type, and Issuer Industry. For Alameda, it seems that kPC1 is better capturing the categorical

Table 9
cKTA results for the climate data set. Each row shows a considered county, while, in the columns, we have the different approximation matrices used for the cKTAs. The first column presents the cKTA calculated using the covariance matrices of the engineered features for the climate data and the rank-one approximation covariance matrices using PC1 as a column vector. The second column presents the cKTA calculated using the covariance matrices of the engineered features for the climate data and the rank-two approximation covariance matrices using PC1 and PC2 as column vectors. Equivalent reasoning applies to the rest of the columns.

Results of centered kernel target alignment - climate dataset

County	PC1	PC1 & PC2	PC1 & PC2 & PC3	kPC1	kPC1 & kPC2	kPC1 & kPC2 & kPC3
Alameda	0.434	**0.764**	**0.767**	0.515	**0.794**	**0.823**
Los Angeles	0.250	0.591	0.582	0.425	**0.733**	**0.842**
Napa	0.511	**0.818**	**0.841**	0.361	**0.819**	**0.920**
San Diego	0.245	0.617	0.606	0.230	0.525	0.628
San Francisco	0.657	**0.741**	**0.741**	0.456	0.601	**0.806**
San Joaquin	**0.740**	**0.756**	**0.800**	0.676	**0.780**	**0.885**
San Luis Obispo	0.528	**0.804**	**0.817**	0.654	**0.821**	**0.951**
Santa Clara	0.483	**0.768**	**0.773**	0.522	**0.898**	**0.925**
Santa Cruz	0.390	**0.745**	**0.736**	0.258	0.653	**0.773**

Table 10
cKTA results for the financial data set. Each row shows a considered county, while, in the columns, we have the different approximation matrices used for the cKTAs. The first column presents the cKTAs calculated using the covariance matrices of the financial data and the rank-one approximation covariance matrices using PC1 as a column vector. The second column presents the cKTAs calculated using the covariance matrices of the financial data set and the rank-two approximation covariance matrices using PC1 and PC2 as column vectors. Equivalent reasoning applies to the rest of the columns.

Results of centered kernel target alignment - financial dataset

County	PC1	PC1 & PC2	PC1 & PC2 & PC3	kPC1	kPC1 & kPC2	kPC1 & kPC2 & kPC3
Alameda	0.159	0.345	0.567	0.115	0.395	**0.788**
Los Angeles	0.372	0.443	0.458	0.493	0.585	**0.730**
Napa	0.255	0.501	0.611	**0.804**	**0.707**	**0.707**
San Diego	0.136	0.435	0.443	**0.765**	**0.845**	**0.837**
San Francisco	0.345	0.377	0.489	0.224	0.428	0.680
San Joaquin	0.476	0.457	0.566	**0.867**	**0.866**	**0.823**
San Luis Obispo	0.345	0.467	0.557	**0.958**	**0.867**	**0.856**
Santa Clara	0.387	0.427	0.655	0.234	0.531	0.649
Santa Cruz	0.329	0.346	0.453	0.135	0.494	**0.782**

variables, while, for Los Angeles and San Francisco is instead kPC2. However, the level of alignments achieved is around 0.4, suggesting a 40% alignment on average. Napa, given the low number of samples, show zero levels of alignments, suggesting the need for a future analysis once more municipal green bonds have been issued in this county. This result is interesting as it suggests that the issuer is influential in the association detected with pollution air quality variation. This can be directly understood as follows, if a municipal issuer of green bonds is responsible for say road and transportation development in the county, they are likely making the green bond issuance and then utilising the raised proceeds for expenditures on items that reduce pollution arising from transportation and road networks, a significant factor in reducing air pollution. A second example as to why such an association may be expected is for muni issuers who may be responsible for energy production, their issuance of green bonds may lead to changes in the manner in which energy production is performed, by incorporating less coal and more solar and wind energy in the counties electricity grid, funded by issuance of green bonds for this purpose. As such, it is rather natural to expect that the issuer category may play an instrumental role in such an association between green bond variables and pollution air quality data.

Regarding, the other leading features, the Muni Issue Type and Muni Offering Type is also important as the appropriate structuring may be influential in the successful issuance of the green bonds and when issuers get this right for the market they are seeking to raise capital, they may make the issuance over subscribed and successful, which will naturally lead to further issuance's and increased funding for the green initiatives for which the raised capital will be deployed, leading to further reductions in air pollution. Lastly, the significance of the issuer industry is clear, since the study focuses directly on pollution as measured via air quality, the issuer industries should be influential factors in the detection of the associations discovered since, the industries particularly related to pollution reduction in airborne particulate matter and gas emissions will most influence the results in this study.

Fig. 18 shows the equivalent analysis for the numerical green bond financial variables. In this plot, compared to the CKTAs of the categorical variable, the direction of the interpretation is reversed, i.e. the more minor the MSE, the better the kPC has captured underlying variations. Overall, it appears that the MSEs of kPC1 are more significant than the ones of kPC2. Napa has a meagre sample size and, therefore, requires more research for a more reliable interpretation. By focusing on kPC1 only, San Luis Obispo, Los Angeles, San Diego, Santa Clara and Santa Cruz appear to have low MSE overall, particularly San Luis Obispo. Alameda, instead, shows higher MSE levels. The variables showing the least MSE for kPC1 across all counties are the amount out, the maturity size, the Bid OAS and the dated date. All financial variables are defined in Table 6. Interestingly, if one refers to Fig. 13 and focuses on the three variables available in the boxplots (amount out, maturity size and Bid OAS), one can observe many variations across the counties, within each variable. kPC1 best captures such variations.

Regarding the kPC2, much smaller MSEs are identified for every county, except for San Joaquin, presenting MSEs of kPC2 much bigger than the first one. This further suggests that kPC1 and kPC2 capture

Table 11
In these tables the results for the PCA-CCA are provided. The left tables refer to the PCA-CCA for the financial and the pollution data, while, the right tables to the financial and climate data instead. The top tables refer to the PC1-CCA model assessments, while, the bottom ones refer to the PCA2-CCA. Each table presents the canonical variate of interest (from 1 to 9), the canonical correlation coefficient ρ^*, the squared canonical correlation ρ^{*^2}, the F-statistic with the Rao's F-approximation test, the two sets of degrees of freedom and the p-value. More details about this statistical part are given in Section 3.3 and Table 1.

Canonical correlation summary financial data set vs pollution data set

PC1						
Main results			F test for canonical correlations (Rao's F approximation)			
CV	ρ^*	ρ^{*^2}	F	df_2	df_1	Pr(>X)
1	0.012	0.109	0.001	81	538.90	1.000
2	0.002	0.044	0.001	64	485.22	1.000
3	0.001	0.031	0.001	49	430.88	1.000
4	0.001	0.031	0.001	36	376.02	0.998
5	0.000	0.000	0.001	25	320.98	0.987
6	0.000	0.000	0.001	16	266.43	0.999
7	0.000	0.000	0.000	9	214.32	1.000
8	0.000	0.000	0.000	4	178.00	1.000
9	0.000	0.000	0.000	1	90.00	0.999

(a) PC1-CCA Model Assessment.

PC2						
Main results			F test for canonical correlations (Rao's F approximation)			
CV	ρ^*	ρ^{*^2}	F	df_2	df_1	Pr(>X)
1	0.006	0.077	0.012	81	538.90	1.000
2	0.005	0.070	0.011	64	485.22	1.000
3	0.002	0.044	0.001	49	430.88	1.000
4	0.002	0.044	0.001	36	376.02	1.000
5	0.001	0.031	0.001	25	320.98	1.000
6	0.000	0.000	0.000	16	266.43	1.000
7	0.000	0.000	0.000	9	214.32	1.000
8	0.000	0.000	0.000	4	178.00	1.000
9	0.000	0.000	0.000	1	90.00	1.000

(b) PC2-CCA Model Assessment.

Canonical correlation summary financial data set vs climate data set

PC1						
Main results			F test for canonical correlations (Rao's F approximation)			
CV	ρ^*	ρ^{*^2}	F	df_2	df_1	Pr(>X)
1	0.009	0.094	0.002	81	538.90	1.000
2	0.005	0.070	0.002	64	485.22	1.000
3	0.003	0.054	0.001	49	430.88	1.000
4	0.001	0.031	0.001	36	376.02	1.000
5	0.000	0.000	0.001	25	320.98	1.000
6	0.000	0.000	0.000	16	266.43	1.000
7	0.000	0.000	0.000	9	214.32	1.000
8	0.000	0.000	0.000	4	178.00	1.000
9	0.000	0.000	0.000	1	90.00	1.000

(c) PC1-CCA Model Assessment.

PC2						
Main results			F test for canonical correlations (Rao's F approximation)			
CV	ρ^*	F	ρ^{*^2}	df_2	df_1	Pr(>X)
1	0.007	0.083	0.002	81	538.90	1.000
2	0.001	0.031	0.002	64	485.22	1.000
3	0.001	0.031	0.016	49	430.88	1.000
4	0.001	0.031	0.001	36	376.02	1.000
5	0.000	0.000	0.000	25	320.98	1.000
6	0.000	0.000	0.000	16	266.43	1.000
7	0.000	0.000	0.000	9	214.32	1.000
8	0.000	0.000	0.000	4	178.00	1.000
9	0.000	0.000	0.000	1	90.00	1.000

(d) PC2-CCA Model Assessment.

(a) kPC1-CV1.

(b) kPC2-CV1.

Fig. 15. Structured coefficients of financial/pollution kPC1 (left) and kPC2 (right) for the first canonical variate. These are presented in Table 13, in the fourth column.

many different underlying variations of the data, leading to different levels of MSEs. Moreover, the highest levels of kPC1 might also be related to a greater captured variation than kPC2, yielding more significant errors.

Table 12
In these tables the results for the kPCA-CCA are provided. The left tables refer to the kPCA-CCA for the financial and the pollution data, while, the right tables to the financial and climate data instead. The top tables refer to the kPC1-CCA model assessments, while, the bottom ones refer to the kPCA2-CCA. Each table presents the canonical variate of interest (from 1 to 9), the canonical correlation coefficient ρ^*, the squared canonical correlation ρ^{*2}, the F-statistic with the Rao's F-approximation test, the two sets of degrees of freedom and the p-value. More details about this statistical part are given in Section 3.3 and Table 1.

Canonical correlation summary financial data set vs pollution data set

kPC1						
Main results			F test for canonical correlations (Rao's F approximation)			
CV	ρ^*	ρ^{*2}	F	df_2	df_1	Pr(>X)
1	**1.000**	**0.999**	853.793	81	6355.2	**< 2.2e − 16**
2	**0.790**	**0.724**	36.309	64	5676.3	**< 2.2e − 16**
3	0.547	0.299	22.203	49	5000.0	**< 2.2e − 16**
4	0.488	0.238	18.984	36	4328.2	**< 2.2e − 16**
5	0.422	0.178	15.388	25	3664.3	**< 2.2e − 16**
6	0.315	0.099	11.039	16	3016.0	**< 2.2e − 16**
7	0.230	0.052	7.697	9	2404.7	**3.134e − 11**
8	0.111	0.012	3.660	4	1978.0	**0.005**
9	0.047	0.002	2.225	1	990.0	0.136

(a) kPC1-CCA Model Assessment.

kPC2						
Main results			F test for canonical correlations (Rao's F approximation)			
CV	ρ^*	ρ^{*2}	F	df_2	df_1	Pr(>X)
1	**1.000**	**0.999**	1082.3	81	6355.2	**< 2.2e − 16**
2	**0.881**	**0.776**	51.524	64	5676.3	**< 2.2e − 16**
3	0.638	0.407	25.810	49	5000.0	**< 2.2e − 16**
4	0.459	0.211	18.292	36	4328.2	**< 2.2e − 16**
5	0.388	0.151	15.998	25	3664.3	**< 2.2e − 16**
6	0.358	0.128	14.149	16	3016.0	**< 2.2e − 16**
7	0.212	0.045	9.369	9	2404.7	**4.097e − 14**
8	0.179	0.032	9.390	4	1978.0	**1.616e − 07**
9	0.069	0.004	4.780	1	990.0	0.029

(b) kPC2-CCA Model Assessment.

Canonical correlation summary financial data set vs climate data set

kPC1						
Main results			F test for canonical correlations (Rao's F approximation)			
CV	ρ^*	ρ^{*2}	F	df_2	df_1	Pr(>X)
1	**0.815**	**0.712**	27.601	81	6355.2	**< 2.2e − 16**
2	0.673	0.453	21.085	64	5676.3	**< 2.2e − 16**
3	0.491	0.241	13.377	49	5000.0	**< 2.2e − 16**
4	0.423	0.179	9.939	36	4328.2	**< 2.2e − 16**
5	0.259	0.067	6.074	25	3664.3	**< 2.2e − 16**
6	0.215	0.046	5.067	16	3016.0	**1.53e − 10**
7	0.165	0.027	3.686	9	2404.7	**0.001**
8	0.073	0.005	1.397	4	1978.0	0.232
9	0.014	0.000	0.202	1	990.0	0.652

(c) kPC1-CCA Model Assessment.

kPC2						
Main results			F test for canonical correlations (Rao's F approximation)			
CV	ρ^*	F	ρ^{*2}	df_2	df_1	Pr(>X)
1	0.664	0.441	22.827	81	6355.2	**< 2.2e − 16**
2	0.572	0.327	18.038	64	5676.3	**< 2.2e − 16**
3	0.515	0.265	14.406	49	5000.0	**< 2.2e − 16**
4	0.374	0.140	10.309	36	4328.2	**< 2.2e − 16**
5	0.347	0.120	8.523	25	3664.3	**< 2.2e − 16**
6	0.242	0.058	5.093	16	3016.0	**1.294e − 10**
7	0.120	0.014	2.288	9	2404.7	0.014
8	0.069	0.004	1.496	4	1978.0	0.200
9	0.001	0.001	1.353	1	990.0	0.244

(d) kPC2-CCA Model Assessment.

(a) kPC1-CV1.

(b) kPC2-CV1.

Fig. 16. Structured coefficients of financial/climate kPC1 (left) and kPC2 (right) for the first canonical variate. These are presented in Table 13, in the fourth column.

6. Discussion and conclusion

Green bonds are distinctive financial instruments that direct funds towards environmentally advantageous initiatives, setting them apart from their conventional bond counterparts. However, assessing the potential for environmental and climate mitigations of this nascent green bond market presents a considerable challenge for investors due to the lack of standardised reporting on environmental impact.

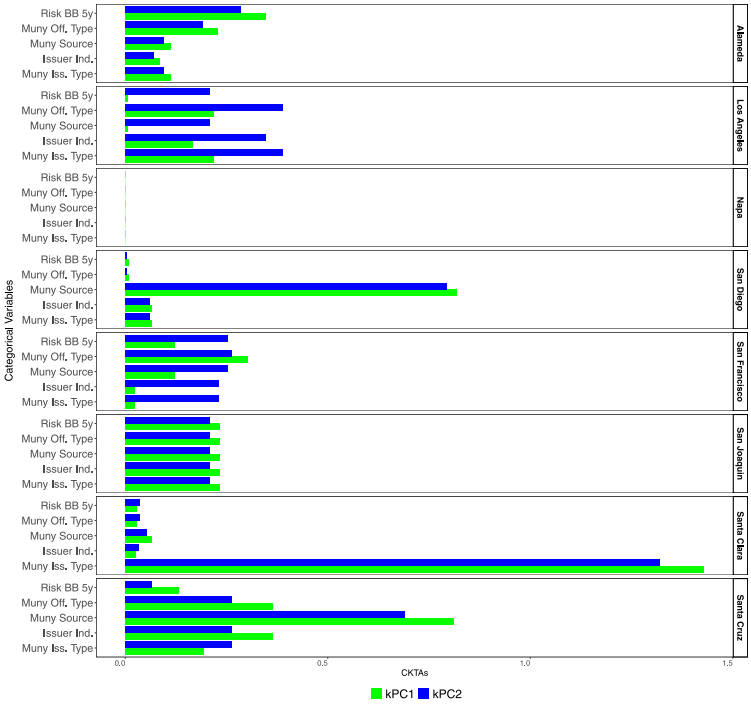

Fig. 17. cKTAs of kPC1 and kPC2 by county in capturing each individual categorical variable. The alignments are computed between the empirical covariance matrix of a vector kPC and the empirical covariance matrix of a vector categorical variable. In the y-axis the categorical variables are given and the x-axis represents the cKTA. The cKTA is comprised in a range between -1 and 1.

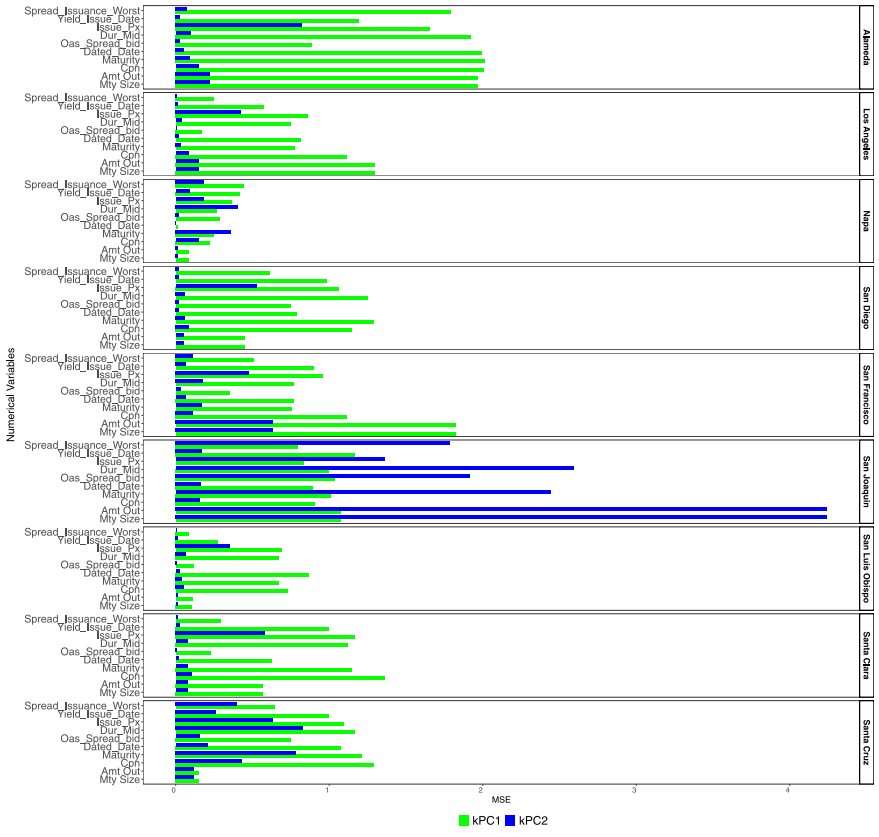

Fig. 18. MSEs of kPC1 and kPC2 by county in capturing each individual numerical variable. The MSE is computed through the reconstructed pre-images obtained kPC1 and kPC1 and kPC2 and then the euclidean distances between the reconstructed data and the original data are computed. In the y-axis the numerical variables are given and the x-axis represents the MSEs..

Table 13
In these tables the results for the Canonical Correlation Analysis are provided. The left table refer to the kPCA-CCA for the financial and pollution data sets, while, the right table to the financial and climate data sets. Further, each table is split with the top panel for the results of kPC1 and the bottom panel for kPC2. In columns there are the structured coefficients and the squared structured coefficients for the first and the second canonical variates, respectively. Note that Figs. 15 and 16 show results of the structured coefficients using helio plots.

kPCA-CCA - Financial data set vs pollution data set

			Structured Coef.		Squared Structured Coef.	
Data set	kPC	County	CV 1	CV 2	CV 1	CV 2
Financial	1	Alameda	**0.717**	0.122	0.514	0.015
Financial	1	Los Angeles	−0.442	0.075	0.195	0.006
Financial	1	Napa	**−0.999**	−0.013	**0.999**	0.000
Financial	1	San Diego	0.117	0.231	0.014	0.053
Financial	1	San Francisco	0.200	0.168	0.040	0.028
Financial	1	San Joaquin	**1.000**	0.000	**1.000**	0.000
Financial	1	San Luis Obispo	**−1.000**	0.000	**1.000**	0.000
Financial	1	Santa Clara	−0.380	−0.646	0.145	0.418
Financial	1	Santa Cruz	−0.521	−0.410	0.271	0.168
Pollution	1	Alameda	**−0.924**	0.138	**0.854**	0.019
Pollution	1	Los Angeles	**0.651**	0.337	0.423	0.113
Pollution	1	Napa	**0.766**	0.222	0.587	0.049
Pollution	1	San Diego	**0.941**	−0.151	**0.885**	0.023
Pollution	1	San Francisco	**0.770**	0.209	0.593	0.044
Pollution	1	San Joaquin	**−0.913**	0.073	**0.834**	0.005
Pollution	1	San Luis Obispo	**0.930**	−0.175	**0.865**	0.031
Pollution	1	Santa Clara	**0.917**	−0.146	**0.841**	0.021
Pollution	1	Santa Cruz	**0.912**	−0.150	**0.831**	0.023
Financial	2	Alameda	**0.861**	−0.113	**0.742**	0.013
Financial	2	Los Angeles	−0.348	−0.019	0.121	0.000
Financial	2	Napa	**0.999**	0.014	**0.999**	0.000
Financial	2	San Diego	**0.848**	0.041	0.022	0.002
Financial	2	San Francisco	**0.849**	0.010	0.062	0.000
Financial	2	San Joaquin	**1.000**	0.000	**1.000**	0.000
Financial	2	San Luis Obispo	**−1.000**	0.001	**1.000**	0.000
Financial	2	Santa Clara	−0.342	0.427	0.117	0.183
Financial	2	Santa Cruz	0.508	0.344	0.258	0.118
Pollution	2	Alameda	**−0.695**	−0.629	0.483	0.395
Pollution	2	Los Angeles	**−0.907**	0.156	**0.823**	0.024
Pollution	2	Napa	**0.883**	−0.290	**0.779**	0.084
Pollution	2	San Diego	**−0.881**	0.319	**0.776**	0.102
Pollution	2	San Francisco	**−0.788**	0.513	0.621	0.263
Pollution	2	San Joaquin	**−0.915**	0.154	**0.837**	0.024
Pollution	2	San Luis Obispo	0.365	**0.838**	0.133	**0.703**
Pollution	2	Santa Clara	−0.401	**−0.682**	0.161	0.466
Pollution	2	Santa Cruz	−0.307	**0.869**	0.094	**0.755**

(a) Canonical Correlation Analysis

kPCA-CCA - Financial data set vs climate data set

			Structured Coef.		Squared Structured Coef.	
Data set	kPC	County	CV 1	CV 2	CV 1	CV 2
Financial	1	Alameda	−0.081	−0.552	0.007	0.305
Financial	1	Los Angeles	0.307	0.368	0.094	0.135
Financial	1	Napa	0.556	0.310	0.309	0.096
Financial	1	San Diego	−0.135	−0.558	0.018	0.311
Financial	1	San Francisco	0.130	−0.308	0.017	0.095
Financial	1	San Joaquin	−0.543	−0.338	0.295	0.114
Financial	1	San Luis Obispo	0.543	0.338	0.295	0.114
Financial	1	Santa Clara	−0.040	0.412	0.002	0.170
Financial	1	Santa Cruz	0.009	0.358	0.000	0.128
Climate	1	Alameda	−0.323	0.175	0.105	0.031
Climate	1	Los Angeles	−0.294	0.382	0.087	0.146
Climate	1	Napa	**0.708**	−0.103	0.502	0.011
Climate	1	San Diego	**0.832**	−0.172	**0.692**	0.030
Climate	1	San Francisco	−0.042	−0.145	0.002	0.021
Climate	1	San Joaquin	**0.733**	0.314	0.537	0.099
Climate	1	San Luis Obispo	0.571	−0.453	0.326	0.206
Climate	1	Santa Clara	0.473	0.229	0.224	0.053
Climate	1	Santa Cruz	0.363	0.082	0.131	0.007
Financial	2	Alameda	0.401	−0.535	0.161	0.286
Financial	2	Los Angeles	−0.082	−0.347	0.007	0.121
Financial	2	Napa	−0.113	**0.780**	0.013	0.608
Financial	2	San Diego	−0.462	0.490	0.213	0.240
Financial	2	San Francisco	−0.222	0.109	0.049	0.012
Financial	2	San Joaquin	−0.139	**0.773**	0.019	0.598
Financial	2	San Luis Obispo	−0.137	**0.773**	0.019	0.597
Financial	2	Santa Clara	−0.258	−0.385	0.067	0.148
Financial	2	Santa Cruz	−0.318	0.414	0.101	0.172
Climate	2	Alameda	0.274	0.258	0.075	0.066
Climate	2	Los Angeles	0.420	0.079	0.177	0.006
Climate	2	Napa	0.303	**0.684**	0.092	0.468
Climate	2	San Diego	−0.430	**−0.769**	0.185	0.591
Climate	2	San Francisco	0.188	−0.264	0.036	0.070
Climate	2	San Joaquin	0.101	**−0.782**	0.010	0.612
Climate	2	San Luis Obispo	−0.579	−0.418	0.335	0.175
Climate	2	Santa Clara	−0.103	0.434	0.011	0.188
Climate	2	Santa Cruz	0.044	0.435	0.002	0.190

(b) Canonical Correlation Analysis

Our research initiative designed a unique set of indicators as a first stage of an ongoing process designed to monitor and address this gap, leveraging financial and environmental data sets and employing sophisticated statistical techniques.

The methodology and experimental design applied in this study facilitated an in-depth and multifaceted exploration of the influence of green bonds on environmental and climate-associated parameters in Californian counties. California, chosen for its abundant data availability and numerous environmental monitoring stations, offered an ideal backdrop for our investigation. Although the positioning of these stations introduced certain complexities, they did not detract from the overall viability of the study. The research focused on key cities in California, incorporating areas within a 50 km radius. This decision imbued the study with a layer of practical realism, as these zones frequently serve as the nucleus for dynamic economic and environmental operations, making them prime areas for the likely tangible effects of green bond issuance.

Integrating three different data sets, pollution, climate, and green bonds, into our research approach, we achieved a multifaceted view of the complex interplay between fiscal incentives and environmental improvement. This multidimensional perspective, together with our rigorous methodology, underscores the potential and importance of green bonds in instigating significant positive environmental change, providing a valuable reference for investors interested in environmentally responsible investment opportunities.

Our study compared Principal Component Analysis-Canonical Correlation Analysis (PCA-CCA) and kernel Principal Component Analysis-Canonical Correlation Analysis (kPCA-CCA). The focus of this comparison was to evaluate the ability of each methodology to capture cross-correlation variability within spatial–temporal multivariate data, evaluating the impact of municipal green bonds as a whole within California by considering pollution and climate as attributes of the desired impact. Central to our research was applying kPCA and CCA to identify cross-correlation within spatial–temporal multivariate data sets. This novel approach allows for the combination of spatial–temporal multivariate data sources, with several recording frequencies, different structure data types, and different recording spatial observation collections. In particular, it offers a unique solution to handling issues related to variable comparability and managing the differential treatment of categorical and numerical variables. This method adopts a progressive

strategy to address the challenges associated with disparate variable types in multivariate data sets. Using kPCA and CCA in conjunction, we could uncover nuanced relationships within the data that would otherwise have been difficult to discern with more conventional analytical techniques.

The kPCA method, an extension of the traditional PCA, effectively deals with non-linearity in the data set by mapping the input into a higher-dimensional feature space. In this high-dimensional space, we can perform CCA, to tease out the complex structures of cross-correlations in the data set that are not immediately apparent. In essence, by harnessing the combined power of kPCA and CCA, our research could innovatively tackle the intricacies of multivariate data sets, provide more reliable results, and offer a more nuanced understanding of the potential impacts of green bonds. Thus, this approach significantly contributes to developing advanced data analysis in sustainable finance and environmental impact assessment.

Although the first PC represented more than 50% of the variance within the pollution and climate data sets, the first three PCs could only detect less than 30% of the data variability in the financial data set. These findings underscore the relative strengths and limitations of PCA, particularly its struggle to effectively capture the non-stationarity nature of these data. In contrast, kPCA-CCA demonstrated more uniform explanatory power across the three kPCs, particularly in high non-stationarity levels. This finding further reinforced the decision to adopt the kPCA-CCA approach in this study. Centred Empirical Kernel Alignment (cKTA) results corroborated the superiority of kPCs in capturing the underlying engineered pollution features compared to PCs.

When analysing climate data, both PCs and kPCs exhibited high cKTAs, indicating efficient capture of variability in climate characteristics across different counties. It is particularly noteworthy that kPCs achieved over 90% alignment in some counties, demonstrating the utility of kPCA-CCA in managing the non-stationarity in the data. kPCs strongly outperformed PCs in the case of the financial data set, with higher levels of alignment achieved for all counties except San Francisco and Santa Clara.

Applying CCA to the PCA results, it was discovered that the canonical correlation was neither high nor statistically significant for any of the selected PCs. This suggests that PCs did not capture the global presence of non-stationarity in the data, pointing out the limitations of using traditional PCA in such a context. In stark contrast, kPCA-CCA revealed high levels of correlation for the first two canonical functions of kPC1 and kPC2, especially in the pollution versus financial data set, revealing the potential power of kPCA in uncovering the relationships between complex data sets. Another critical observation made during the research was the lower correlation between modes of variations extracted from financial data and climate data compared to financial data and pollution data. It may imply that the impacts of green bonds on climate variables are less immediate and more long-term, making them less observable in the immediate term. Notably, the study also emphasised the importance of squared canonical correlation. Despite the rate of canonical correlation decreasing slower than that of squared canonical correlation, the study highlighted the possibility of a low shared variance between synthetic canonical variates, even if the correlation is maximised. This indicates that if the variance shared between synthetic canonical variates is low, the corresponding pair of canonical functions will not carry significant information.

This research unravels the intricate relationships between the issuance of green bonds and their environmental and climatic impacts, focusing on California. Our approach employed multidimensional data analysis, rigorous data preparation procedures, and advanced analytical methodologies, such as kPCA and CCA, enhanced by hyperparameter learning. This comprehensive analytical approach yielded significant insights into the complex dynamics interconnecting green bond issuance and environmental impacts.

Our research unearthed some notable findings when we applied the innovative kPCA-CCA methodology to analyse municipal financial data associated with green bonds and pollution data from nine California counties. A clear and interpretable correlation emerged from the analysis, directly related to green bond issuance. This correlation provides tangible evidence of these financial instruments' impact on promoting environmental improvements. Furthermore, our study highlighted specific patterns at the county level, revealing, for example, a negative correlation between financial and pollution variables in counties such as Alameda and San Joaquin. These results stress the nuanced locality-specific dynamics interweaving green bond issuance with environmental outcomes, highlighting the importance of localised in-depth analyses. Such a negative correlation, also found in San Francisco, San Diego and Napa, can be interpreted when it is put in the context of CCA structured coefficients analysis. The results show a positive impact within these counties, directly interpretable from the developed methodology. Furthermore, different kPCs captured different variation frequencies, suggesting that this methodology is the right road for such a big purpose.

The outcome of such research would improve the transparency of the green bond market and reinforce investor confidence in green bonds. This is particularly important given green bonds' critical role in facilitating the economic transition required to achieve the targets established in the Paris Agreement.

The insights from this research have substantial implications for decision-making processes related to green bonds. With the robust kPCA-CCA methodology, stakeholders can obtain detailed and nuanced insights into the relationships between the financial aspects of green bonds and pollution or climate variables. These insights can then guide the creation and implementation of green bond strategies that truly advance environmental sustainability.

Our research emphasises the central role of green bonds in driving environmental progress and minimising climate change. Advanced methodologies like kPCA-CCA can lead to more informed decision-making and strategic development, reinforcing the role of green bonds as integral financial tools for promoting a sustainable future. Nonetheless, these relationships' complexity and multiple facets necessitate ongoing research, especially over extended timescales, to fully understand the long-term impacts of green bonds on our climate.

Lastly, this research reveals essential insights into the complex relationships between the financial variables of green bonds and pollution/climate data. Our novel analytical approach, involving PCA, kPCA, and CCA, enabled us to dissect these relationships in-depth, revealing both the strengths and limitations of each methodology and thus contributing to a more comprehensive understanding of the impacts and potential of green bonds.

CRediT authorship contribution statement

Marta Campi: Data curation, Formal analysis, Methodology, Software, Visualization, Writing – original draft, Writing – review & editing, Conceptualization. **Gareth W. Peters:** Conceptualization, Investigation, Methodology, Project administration, Supervision, Writing – original draft, Writing – review & editing. **Kylie-Anne Richards:** Data curation, Formal analysis, Investigation, Software, Supervision, Visualization, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.fraope.2024.100113.

References

[1] S. Solomon, Climate Change 2007-The Physical Science Basis: Working Group I Contribution to the Fourth Assessment Report of the IPCC, vol. 4, Cambridge University Press, 2007.
[2] T.L. Root, J.T. Price, K.R. Hall, S.H. Schneider, C. Rosenzweig, J.A. Pounds, Fingerprints of global warming on wild animals and plants, Nature 421 (6918) (2003) 57.
[3] P.M. Cox, R.A. Betts, C.D. Jones, S.A. Spall, I.J. Totterdell, Acceleration of global warming due to carbon-cycle feedbacks in a coupled climate model, Nature 408 (6809) (2000) 184.
[4] K. Protocol, Kyoto, UNFCCC, COP3, Japan, 1997.
[5] S. Oberthür, H.E. Ott, The Kyoto Protocol: International Climate Policy for the 21st Century, Springer Science & Business Media, 1999.
[6] K. Protocol, United Nations framework convention on climate change, Kyoto Protocol, Kyoto 19 (1997).
[7] N. Stern, et al., What is the economics of climate change? World Econ.-Henley Thames 7 (2) (2006) 1.
[8] M.C. Fuller, S.C. Portis, D.M. Kammen, Toward a low-carbon economy: municipal financing for energy efficiency and solar power, Environ. Sci. Policy Sustain. Dev. 51 (1) (2009) 22–33.
[9] E. Campiglio, Beyond carbon pricing: The role of banking and monetary policy in financing the transition to a low-carbon economy, Ecol. Econom. 121 (2016) 220–230.
[10] H. Gujba, S. Thorne, Y. Mulugetta, K. Rai, Y. Sokona, Financing low carbon energy access in Africa, Energy Policy 47 (2012) 71–78.
[11] J.-q. Bao, Y. Miao, F. Chen, Low carbon economy: Revolution in the way of human economic development, China Ind. Econ. 4 (2008) (2008) 017.
[12] K. Shimada, Y. Tanaka, K. Gomi, Y. Matsuoka, Developing a long-term local society design methodology towards a low-carbon economy: An application to Shiga Prefecture in Japan, Energy Policy 35 (9) (2007) 4688–4703.
[13] A.P. Kinzig, D.M. Kammen, National trajectories of carbon emissions: analysis of proposals to foster the transition to low-carbon economies, Global Environ. Change 8 (3) (1998) 183–208.
[14] N. Stern, Stern review report on the economics of climate change, 2006.
[15] S. Griffith-Jones, J. Tyson, The European investment bank: Lessons for developing countries, 2013, WIDER Working Paper 2013/019.
[16] International Capital Market Association, et al., Green Bond Principles: Voluntary Process Guidelines for Issuing Green Bonds, International Capital Market Association, Zürich, 2018.
[17] G. Peters, R. Zhu, G. Tzougas, G. Rabitti, I. Yusuf, The role and significance of green bonds in funding transition to a low carbon economy: A case study forecasting portfolios of green bond instrument returns, 2022, Available at SSRN 4299196.
[18] International Capital Market Association, et al., Green bond principles, 2014, Retrieved from International Capital Market Association website: http://www.icmagroup.org/Regulatory-Policy-and-Market-Practice/green-bonds/green-bond-principles.
[19] C. Wang, P. Liu, H. Ibrahim, R. Yuan, The temporal and spatial evolution of green finance and carbon emissions in the Pearl River Delta region: An analysis of impact pathways, J. Clean. Prod. (2024) 141428.
[20] L. Gao, K. Guo, X. Wei, Dynamic relationship between green bonds and major financial asset markets from the perspective of climate change, Front. Environ. Sci. 10 (2023) 1109796.
[21] A. Dan, A. Tiron-Tudor, The determinants of green bond issuance in the European union, J. Risk Financ. Manag. 14 (9) (2021) 446.
[22] S. Yi, Y. Zhou, J. Zhang, Q. Li, Y. Liu, Y. Guo, Y. Chen, Spatial-temporal evolution and motivation of ecological vulnerability based on RSEI and GEE in the Jianghan Plain from 2000 to 2020, Front. Environ. Sci. 11 (2023) 1191532.
[23] Y. Zhao, N. Zhao, R. Lyu, The dynamic coupling and spatio-temporal differentiation of green finance and industrial green transformation: Evidence from China regions, Heliyon 9 (12) (2023).
[24] Y. Zhang, J. She, X. Long, M. Zhang, Spatio-temporal evolution and driving factors of eco-environmental quality based on RSEI in Chang-Zhu-Tan metropolitan circle, central China, Ecol. Indic. 144 (2022) 109436.
[25] M. Czech, M. Hadaś-Dyduch, B. Puszer, Effectiveness of green bonds in selected CEE countries: Analysis of similarities, Risks 11 (12) (2023) 214.
[26] C.-C. Lee, F. Liu, J. Shi, What impacts do green bonds have on carbon emissions and how? A dynamic spatial perspective in China, Environ. Sci. Pollut. Res. 30 (55) (2023) 117981–117997.
[27] American Lung Association, State of the Air, American Lung Association, 2023.
[28] L. Fisher, S. Ziaja, Statewide summary report, Calif. Fourth Clim. Assess. (2018).
[29] B. Schölkopf, A. Smola, K.-R. Müller, Nonlinear component analysis as a kernel eigenvalue problem, Neural Comput. 10 (5) (1998) 1299–1319.
[30] W.K. Härdle, L. Simar, Canonical correlation analysis, in: Applied Multivariate Statistical Analysis, Springer, 2015, pp. 443–454.
[31] H. Hotelling, The most predictable criterion, J. Educ. Psychol. 26 (2) (1935) 139.
[32] S. Meng, C. Tong, T. Lan, H. Yu, Canonical correlation analysis-based explicit relation discovery for statistical process monitoring, J. Franklin Inst. 357 (8) (2020) 5004–5018.
[33] I.T. Joliffe, B. Morgan, Principal component analysis and exploratory factor analysis, Stat. Methods Med. Res. 1 (1) (1992) 69–95.
[34] V. Vapnik, The Nature of Statistical Learning Theory, Springer science & business media, 1999.
[35] F.R. Bach, M.I. Jordan, Kernel independent component analysis, J. Mach. Learn. Res. 3 (Jul) (2002) 1–48.
[36] D.R. Hardoon, S. Szedmak, J. Shawe-Taylor, Canonical correlation analysis: An overview with application to learning methods, Neural Comput. 16 (12) (2004) 2639–2664.
[37] K. Yoshida, J. Yoshimoto, K. Doya, Sparse kernel canonical correlation analysis for discovery of nonlinear interactions in high-dimensional data, BMC Bioinformatics 18 (2017) 1–11.
[38] W. Wang, K. Livescu, Large-scale approximate kernel canonical correlation analysis, 2015, arXiv preprint arXiv:1511.04773.
[39] D. Lopez-Paz, S. Sra, A. Smola, Z. Ghahramani, B. Schölkopf, Randomized nonlinear component analysis, in: International Conference on Machine Learning, PMLR, 2014, pp. 1359–1367.
[40] V. Uurtio, S. Bhadra, J. Rousu, Sparse non-linear cca through hilbert-schmidt independence criterion, in: 2018 IEEE International Conference on Data Mining, ICDM, IEEE, 2018, pp. 1278–1283.
[41] V. Uurtio, S. Bhadra, J. Rousu, Large-scale sparse kernel canonical correlation analysis, in: International Conference on Machine Learning, PMLR, 2019, pp. 6383–6391.
[42] J. Xu, W. Li, X. Liu, D. Zhang, J. Liu, J. Han, Deep embedded complementary and interactive information for multi-view classification, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 34, No. 04, 2020, pp. 6494–6501.
[43] N.Y. Bilenko, J.L. Gallant, Pyrcca: regularized kernel canonical correlation analysis in python and its applications to neuroimaging, Front. Neuroinform. 10 (2016) 49.
[44] P. Honeine, C. Richard, A closed-form solution for the pre-image problem in kernel-based machines, J. Signal Process. Syst. 65 (3) (2011) 289–299.
[45] G.H. Bakır, J. Weston, B. Schölkopf, Learning to find pre-images, Adv. Neural Inf. Process. Syst. 16 (2004) 449–456.
[46] C. Saunders, A. Gammerman, V. Vovk, Ridge regression learning algorithm in dual variables, 1998.
[47] P.V. Dattalo, A demonstration of canonical correlation analysis with orthogonal rotation to facilitate interpretation, 2014.
[48] A. Sherry, R.K. Henson, Conducting and interpreting canonical correlation analysis in personality research: A user-friendly primer, J. Pers. Assess. 84 (1) (2005) 37–48.
[49] M.S. Levine, Canonical Analysis and Factor Comparison, (no. 6) Sage, 1977.
[50] C.R. Rao, C.R. Rao, M. Statistiker, C.R. Rao, C.R. Rao, Linear Statistical Inference and Its Applications, vol. 2, Wiley New York, 1973.
[51] S. Lee, J. Choi, Z. Fang, F.D. Bowman, Longitudinal canonical correlation analysis, J. R. Stat. Soc. Ser. C. Appl. Stat. 72 (3) (2023) 587–607, http://dx.doi.org/10.1093/jrsssc/qlad022.
[52] T. Courville, B. Thompson, Use of structure coefficients in published multiple regression articles: β is not enough, Educ. Psychol. Meas. 61 (2) (2001) 229–248.
[53] R.K. Henson, The Logic and Interpretation of Structure Coefficients in Multivariate General Linear Model Analyses, ERIC, 2002.
[54] A. Alfons, C. Croux, P. Filzmoser, Robust maximum association estimators, J. Amer. Statist. Assoc. 112 (517) (2017) 436–445.
[55] P. Molnár, High-low range in GARCH models of stock return volatility, Appl. Econ. 48 (51) (2016) 4977–4991.
[56] P. Cerda, G. Varoquaux, B. Kégl, Similarity encoding for learning with dirty categorical variables, Mach. Learn. 107 (8–10) (2018) 1477–1494.
[57] C. Cortes, M. Mohri, A. Rostamizadeh, Two-stage learning kernel algorithms, 2010.