SHADES OF GREEN: UNVEILING THE IMPACT OF MUNICIPAL GREEN BONDS ON THE ENVIRONMENT

Marta Campi
Institut Pasteur
marta.campi@pasteur.fr

Gareth W. Peters
University of Santa Barbara
garethpeters@ucsb.edu

Kylie-Anne Richards
University of Technology Sydney
kylie-anne.richards@uts.edu.au

July 18, 2023

ABSTRACT

Green bonds allocate proceeds towards environmentally beneficial projects, distinguishing themselves from traditional bonds. 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 advanced statistical techniques. A novel method using kernel Principal Component Analysis (kPCA) and Canonical Correlation Analysis (CCA) is applied to detect cross-correlation in multivariate datasets. This approach innovatively 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 municipal financial data and pollution data from nine California counties. The results show a clear and interpretable correlation directly linked to the amount of green bond issuance, underscoring the tangible impact of these financial instruments on environmental outcomes. However, with the weak cross-correlation observed between climate and financial data sets, further research is recommended over a broader timescale. Such research will confirm the direct effect of green bonds on climate change reduction, enhancing market transparency, and bolstering confidence in green bonds as a crucial tool for the economic transition required by the Paris Agreement.

***K*eywords** Green Bonds, Kernel Principal Component Analysis, Canonical Correlation Analysis

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], [2], [3], [4], [5], [6]. Despite the multifaceted nature of these challenges, financial markets have emerged as instrumental mechanisms in steering the transition towards low-carbon economies [7], [8], [9], [10], [11], [12], [13] and [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, 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 toward 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.

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. Despite the lapse of 16 years since this significant event, the quest for a universally accepted definition of a green bond, extending beyond its taxonomy, remains a topic of active discussion amongst policymakers and market leaders.

In order to provide structure to this growing market, a consortium of investment banks, including Bank of America Merrill Lynch, Citi, Crédit 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 [17]. The GBP was last updated in June 2021, with an additional appendix added in June 2022.

A prominent issue relates to the specification of green projects within bond prospectuses. Several of these 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.

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.

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.

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 to promoting improved environmental outcomes.

A significant gap exists in the academic literature and practical applications of the green bond market: the lack of quantitative environmental impact assessment tools for sovereign, sub-sovereign (including municipal), and state bonds. Therefore, an essential aspect of our research will be 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.

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 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 toward 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 differ-

ent 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 [18]. California also faces serious climate change-related challenges, such as frequent wildfires, persistent droughts, and rising sea levels [19]. This context underscores the pressing need for environmental risk monitoring in the state.

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 ad hoc features. 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: (1) the pollution data has been constructed by downloading information from the US Environmental Protection Agency website[1]; (2) the 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) the 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 elucidate the environmental implications of municipal green bonds. Numerous aspects must be considered. First, the procedure has to deal with the non-stationary and non-linear nature of the data sets. Second, the collected data sets have different data sources, and such a challenge must be efficiently addressed. Third, the environmental impact of green bonds is defined as the statistical association between the modes of variation of the data sets. In this way, accurate correlations can be detected without noisy information polluting the final results. 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 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[20]) and Canonical Correlation Analysis (CCA, [21, 22]).

kPCA is a widely used machine learning technique corresponding to the non-linear version of standard principal component analysis (PCA, see [23]) 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 ([20], [24]) whose idea is to map the existing data set into a new space, called the feature space, where linear algorithms are applied again. 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 nonstationarity.

In this work, the kPCs are the input of the CCA, which then determines the association between them. We remark that 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 sets, in contrast to the new variables of PCA describing the internal variability within one data set. The novelty of this work will be to use the CCA to identify cross-correlation over the robust kPCs extracted by the different data sets. There are several existing methods extending CCA to more robust settings dealing with non-linear setup by kernel methods [25, 26, 27, 28, 29, 30, 31, 32, 33]. These developed methodologies have several advantages, as, for example, kernelizing the CCA problem or applying a deep neural network include the ability to reveal non-linear and non-stationary multivariate relationships in the data. However, the scope of our technique is different from a methodological point of view. These methods consider two input data sets and apply a non-linear and non-stationary transformation to perform CCA or a robust version of CCA on the newly transformed data. Our purpose, instead, is intrinsically different. We will consider several multivariate input data sets, for example, several data sets collecting pollution variables and several data sets describing financial variables. We said several since we will consider one multivariate pair per county, i.e. one multivariate data set for Alameda, one for San Francisco, one for San Diego,

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

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

etc. We will then apply kPCA to extract leading variations from each of the two input data sets and engineer a new multivariate data set containing, for example, the first kernel principal component extracted by the pollution data set and the first kernel principal component extracted by the financial data set[3]. We will repeat such an exercise for every county. At that point, we will have a unique pair of multivariate data, where the attributes are the first kernel principal components of pollution data for every county in one data set and, in the second data set, the first kernel principal components of financial data for every county. Therefore, our analysis is, in a certain way, a multi-multivariate analysis. The reasoning for doing so is that if we look for associations, or correlations, between each county, we will lose relevant cross-correlation information that naturally exists across the counties. The green financial market in California (as everywhere else) is acting in unison; therefore, we must consider what is happening within and between every county. The same idea applies to pollution, as well as to climate. Hence, our innovation lies in revisiting kPCA and CCA and formulating a method for kPCA-CCA considering multiple multivariate data sets.

There is a further catch at this point that must be highlighted. The defined kPCs, when extracted from different counties (this is regardless of the underlying analysed data set), live in a space which is unique to the original data timestamps as well as spatial component (given the different counties considered and, for pollution and climate, the different monitors within the several counties). If fed to a multivariate method as CCA altogether, these kPCs will need to be evaluated through an out-of-sample technique onto a new mesh, common across each county. This further allows to have input data set recorded at different frequencies as well as different spatial locations. Moreover, by using the kernel trick, we are able to extend such a method to variables which are categorical by the use of the Jaccard kernel and efficiently incorporate the information carried by this type of variables. This is highly precious, since, in practice, the treatment of categorical variables is often a burden in multivariate settings. In such a way, we develop a multi-modality kPCA. Our methodology takes these facts into account and develop a novel procedure for mult-multivariate data with different data structures.

We claim that if there is an association between green bond variables and pollution variables (for example), we can interpret such an association as the presence of an impact. 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.

1.1 Contributions, Notation and Organisation of the Paper

The work in this study offers essential contributions, which can be categorised into conceptual, methodological, and data application contributions, as detailed below.

- Our first conceptual contribution provides a quantitative definition of the environmental impact of a green bond, a highly debated topic in the green finance community. We also introduce reliable statistical indicators, specifically the kernel Principal Components (kPCs), that capture variability and highlight leading factors to measure this impact. This approach is a notable requirement in financial markets. Our research shows that these kPCs can detect diversified information based on the data set and the kPC number.
- The second contribution of this work is the combination of the kPCA and the CCA in a novel way considering multiple multivariate pairs of data sets to account for the global green financial market effects as well as the global pollution in the US State of California. The goal is to detect the cross-correlation amongst multiple data sets in non-stationary settings. We show that this approach strongly empowers the desired findings with

[3]We will apply the same reasoning to other kernel principal components, i.e. the second one.

respect to traditional linear PCA combined with CCA. Therefore, it acts as a robust version of CCA for non-linear and non-stationary multi-multivariate data sets.

- Our approach to treating different data sources, including numerical and categorical variables, provides a contribution to data application. This issue often occurs in kernel methods, and we address it by using the Jaccard distance. We incorporate the contribution of the categorical data through the Jaccard kernel, allowing for multi-modality kPCA. This method is instrumental in the analysis of environmental and financial data.
- Another significant contribution to data applications is our development of specific variables and features that are needed to accurately identify the environmental impact over time and space. Given the variations in observational timestamps and spatial recording monitors in the data sets, such feature engineering is essential. It contains pertinent information that encapsulates variability and sheds light on the behaviour of the underlying data.
- A key finding of our study is the substantial and interpretable correlation identified when applying our methodology to green bond municipal financial data and pollution data in nine counties in California. These results are directly related to the volume of green bond issuance, emphasising the real-world environmental impact of these financial instruments. Such investigations can confirm the direct influence of green bonds on climate change reduction, improving market transparency, and increasing faith in green bonds.

Throughout the experimental section, we will be using three different data sets, being the pollution, the climate and the financial data sets that will be introduced at that point. For each of this data sets, we will perform ad hoc data cleaning and wrangling procedures, as well as extract and engineer different sets of features. At this stage, we introduce the notation, required for the understanding of the developed method.

We will denote the three constructed data sets as $\mathbf{X}^1_{N_1\times D_1}$, $\mathbf{X}^2_{N_2\times D_2}$ and $\mathbf{X}^3_{N_3\times D_3}$, where the first one represents the pollution data set, the second one the climate data set and the third one the financial green bond data set, respectively. The first two data sets collect attributes related to pollution or climate (accurately described in section 4) from different pollution or climate stations which are selected according to their distance from the considered counties in California. Therefore, by considering first one $\mathbf{X}^1_{N_1\times D_1}$, the index $N_1 = T_1 \times S$, where T_1 corresponds to the number of timestamps collected for every s locations and S the total number of locations, i.e. the counties in California. D_1 instead represent the signals that have been observed, i.e. carbon dioxide, air quality etc., whereas in the case of $\mathbf{X}^2_{N_2\times D_2}$ the sample notation applies, with $N_2 = T_2 \times S$ and D_2 that represents the observed climate signals as total precipitation, temperature, etc. Regarding $\mathbf{X}^3_{N_3\times D_3}$, which the financial green bonds data set, D_3 represents the observed variables for the green bonds, i.e. coupon, maturity, industry of the green bond, etc. while this time N_3 represents the collected green bonds issued per county but there is no time component associated to it. The set of PCs and KPCs will be extracted by splitting the data set according to each considered county in California, hence we will have $\mathbf{X}^1_{N_1\times D_1} = [\mathbf{x}^{1,1}_{T_1\times D_1}, \mathbf{x}^{1,2}_{T_1\times D_1}, \ldots, \mathbf{x}^{1,S}_{T_1\times D_1}]$, where T_1 is the number of timestamps per county and the upper indices indicate the first data set (pollution) and the number of counties S, respectively. This reasoning is applied to the second data set as well, i.e. the climate one. In the case of the financial data we will have $\mathbf{X}^3_{N_3\times D_3} = [\mathbf{x}^{3,1}_{n_1\times D_3}, \mathbf{x}^{3,2}_{n_2\times D_3}, \ldots, \mathbf{x}^{3,S}_{n_S\times D_3}]$ where n_s for $s = 1, \ldots, S$ is the number of issued and active green bonds per county.

The paper is organised as follows: first, a section reviewing the PCA, the kPCA and the CCA is provided. This includes the review of the out-of-sample and pre-image problems for the Kpca and the introduction of the required notations for each method. Second, a methodology section is presented, with the introduction of the novelty and a toy example showing the methodological contribution. Afterwards, the results for the data and the experiments, are presented. Finally, a discussion of the paper and the conclusions are provided.

2 Spatial-Temporal Methods to Assess Green Bond Disbursements: PCA, kPCA and CCA

This section 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 3 extract PCs and kPCs on the three datasets split by counties, and this will be presented and explained within section 3. For simplicity and without any loss of generality, we review the following methods by referring to a general input matrix $\mathbf{X}_{N\times D}$ and build the utilities required in the paper based on such a matrix.

2.1 PCA for Linear Variability of One Data set

In this section, we firstly introduce the notation required to develop the PCA framework, which will then be extended to its non-linear version in the following subsection. Consider $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N \in \mathcal{X} \subset \mathbb{R}^D$ as our observed input space that we deem as non-linearly transformed $\mathbf{x}_n = [x_{n,1}, \ldots, x_{n,D}]_{1\times D}$. $\mathbf{X}$ the $N \times D$ matrix such that

$$\begin{bmatrix} \mathbf{x}_1 \\ \vdots \\ \mathbf{x}_N \end{bmatrix}_{N\times D} \tag{1}$$

where N represents the number of sample and D the number of variables. We assume that the column of $\mathbf{X}$ have mean zero, i.e. they are centred. The goal of PCA is to identify the most meaningful unit length basis to re-express a given dataset $\mathbf{X}$. Unit based basis means that the length of the vector is 1. We also assume that the new basis is orthogonal, then with the unit length assumption, orthonormality is implied. Therefore, PCA looks for the given projection of the observation data

$$\mathbf{X}_{N\times D}\mathbf{W}_{d\times d} = \mathbf{L}_{N\times d} \tag{2}$$

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 provides by construction $\mathbf{W}^\intercal\mathbf{W} = \mathbb{I}_d$ and express rows of $\mathbf{L}$. The idea behind such a re-expression can be interpreted as the PCA lowering the redundancy in the dataset by removing the linear dependencies. This can be 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 $\mathbf{S}_X = \mathbf{X}^\intercal\mathbf{X}$. In mathematical form

$$\mathbf{S}_{X_{D\times D}} = \begin{bmatrix} \mathbf{x}_{1,1} & \ldots & \mathbf{x}_{1,N} \\ \vdots & \ddots & \vdots \\ \mathbf{x}_{D,1} & \ldots & \mathbf{x}_{D,N} \end{bmatrix}_{D\times N} \begin{bmatrix} \mathbf{x}_{1,1} & \ldots & \mathbf{x}_{1,D} \\ \vdots & \ddots & \vdots \\ \mathbf{x}_{N,1} & \ldots & \mathbf{x}_{N,D} \end{bmatrix}_{N\times D} \tag{3}$$

We seek a linear combination of Equation (2) 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

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

for $\mathbf{\Lambda}_{d\times d}$ 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

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

It is possible to observe that $\mathbf{W}$ is a matrix in which columns are eigenvectors of $\mathbf{S}_X$, whereas $\mathbf{\Lambda}$ 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 as it is an equivalent problem to maximising the variance of $\mathbf{L}$. The final goal is to find the estimates of $\mathbf{W}$ and $\mathbf{L}$ which minimises the sum of squares

$$\epsilon = \mathbf{X} - \mathbf{L}\mathbf{W}^\intercal$$

both for $\epsilon^\intercal\epsilon$ and $\epsilon\epsilon^\intercal$. Assuming that the residuals have homogeneous covariance matrix, i.e. $\epsilon^\intercal\epsilon = \sigma^2\mathbb{I}_d$, we have

$$Q(\mathbf{W}, \mathbf{Z}) = \epsilon^\intercal\epsilon = \sigma^2\mathbb{I}_d = (\mathbf{X} - \mathbf{L}\mathbf{W}^\intercal)^\intercal(\mathbf{X} - \mathbf{L}\mathbf{W}^\intercal) = \mathbf{X}^\intercal\mathbf{X} + \mathbf{W}\mathbf{L}^\intercal\mathbf{L}\mathbf{W} - \mathbf{W}\mathbf{L}^\intercal\mathbf{X} - \mathbf{X}\mathbf{L}\mathbf{W}^\intercal \tag{4}$$

Since both $\mathbf{W}$ and $\mathbf{L}$ are treated as parameters to be estimated, we minimise Equation (4) by computing the partial derivatives of the function Q with respect to $\mathbf{W}$ and $\mathbf{L}$ and setting them to zero

$$\frac{\partial Q}{\partial \mathbf{W}} = -2\mathbf{X}^\intercal\mathbf{L} + 2\mathbf{W}\mathbf{L}^\intercal\mathbf{L} = 0$$

and since by construction $\mathbf{X}\mathbf{W} = \mathbf{L}$, one then obtains

$$\mathbf{X}\mathbf{X}^\intercal\mathbf{L} = \mathbf{X}\mathbf{W}\mathbf{L}^\intercal\mathbf{L} \implies \mathbf{X}\mathbf{X}^\intercal\mathbf{L} = \mathbf{L}\mathbf{L}^\intercal\mathbf{L}$$

As we are looking for uncorrelated explanatory variables, for $\mathbf{\Lambda} = \mathbf{L}^\intercal\mathbf{L}$, we get

$$\mathbf{X}\mathbf{X}^\intercal\mathbf{L} = \mathbf{L}\mathbf{\Lambda}$$

which shows that $\mathbf{L}$ and $\mathbf{\Lambda}$ are eigenvectors and eigenvalues of $\mathbf{X}\mathbf{X}^\intercal$. Furthermore, differentiating Q with respect to $\mathbf{L}$ gives

$$\frac{\partial Q}{\partial \mathbf{L}} = -2\mathbf{X}\mathbf{W} + 2\mathbf{L}\mathbf{W}^\intercal\mathbf{W} = 0$$

By applying $\mathbf{X}\mathbf{W} = \mathbf{L}$, one obtains

$$\mathbf{X}^\intercal\mathbf{X}\mathbf{W} = \mathbf{W}\mathbf{\Lambda}$$

that shows that $\mathbf{W}$ and $\mathbf{\Lambda}$ are eigenvectors and eigenvalues of the covariance matrix of $\mathbf{X}$, i.e. $\mathbf{S}_X = \mathbf{X}^\intercal\mathbf{X}$, respectively.

2.2 kPCA for Non-Linear Variability of One Data set

2.2.1 Background and Main Objectives of the kPCA

The kPCA is employed to detect nonlinear and non-stationary features characterising each of the data sets. We will proceed as above by considering the general input matrix $\mathbf{X}_{N\times D}$ and build the discussion upon it. Assume now $\phi : \mathcal{X} \to \mathcal{F}$, a non-linearly mapping from the observed input space to the linear feature space $\mathcal{F} \subset \mathbb{R}^P$ such that $\phi(\mathbf{x}_n) = [\phi_1(\mathbf{x}_1), \ldots, \phi_P(\mathbf{x}_n)]_{1\times P}$. Remark that, if this mapping was linear, then the PCA would efficiently work. We can write $\phi(\mathbf{x}_n) = \phi_n \in \mathcal{F}$ which will represent the n-th sample out of N samples from the feature space $\mathcal{F}$ and is a P-dimensional vector. $\mathbf{\Phi}$ being an $N \times P$ matrix such that

$$\mathbf{\Phi} = \phi(\mathbf{X}) = \begin{bmatrix} \phi(\mathbf{x}_1) \\ \vdots \\ \phi(\mathbf{x}_N) \end{bmatrix}_{N\times P} = \begin{bmatrix} \phi_1(\mathbf{x}_1) & \ldots & \phi_P(\mathbf{x}_1) \\ \vdots & \ddots & \vdots \\ \phi_1(\mathbf{x}_N) & \ldots & \phi_P(\mathbf{x}_N) \end{bmatrix}_{N\times P} \tag{5}$$

It represents the sample matrix within the feature space. Denote $\mathbf{C}$ as the $P \times P$ positive definite covariance matrix of $\mathbf{\Phi}$ such that

$$\mathbf{C}_{P\times P} = \mathbf{\Phi}^\intercal\mathbf{\Phi} = \begin{bmatrix} \phi_1(\mathbf{x}_1) & \ldots & \phi_1(\mathbf{x}_N) \\ \vdots & \ddots & \vdots \\ \phi_P(\mathbf{x}_1) & \ldots & \phi_P(\mathbf{x}_N) \end{bmatrix}_{P\times N} \begin{bmatrix} \phi_1(\mathbf{x}_1) & \ldots & \phi_P(\mathbf{x}_1) \\ \vdots & \ddots & \vdots \\ \phi_1(\mathbf{x}_N) & \ldots & \phi_P(\mathbf{x}_N) \end{bmatrix}_{N\times P} \tag{6}$$

and one cell of the matrix $\mathbf{C}$ at ith row and jth column is given as

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

and represents covariance between ith feature function ϕ_i and jth feature function ϕ_j within the feature space $\mathcal{F}$ for $i, j = 1, \ldots, P$ across all available samples. Note the differences between the matrices given in Equation 3 and Equation (6). The difference is indeed the presence of a non-linear map defining the covariance matrix of the feature space which required a non-linear method. Now, denote $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ the kernel function that defines the inner product in the feature space $\mathcal{F}$ (here being a dot product that could be generalised) that is given as

$$k(\mathbf{x}_n, \mathbf{x}_m) = \phi(\mathbf{x}_n)\phi(\mathbf{x}_m)^\intercal = [\phi_1(\mathbf{x}_n), \ldots, \phi_P(\mathbf{x}_n)] \begin{bmatrix} \phi_1(\mathbf{x}_m) \\ \vdots \\ \phi_P(\mathbf{x}_m) \end{bmatrix} = \sum_{p=1}^{P} \phi_p(\mathbf{x}_n)\phi_p(\mathbf{x}_m) \tag{7}$$

for $n, m = 1, \ldots, N$. It is an inner product between the samples in the feature space since we can write $\phi(\mathbf{x}_n) = \phi_n \in \mathcal{F}$ is a P dimensional vector within the feature space. Define $\mathbf{K}$ the $N \times N$ Gram Matrix such that

$$\mathbf{K}_{N\times N} = \phi(\mathbf{X})\phi(\mathbf{X})^\intercal = \begin{bmatrix} \phi_1(\mathbf{x}_1) & \ldots & \phi_P(\mathbf{x}_1) \\ \vdots & \ddots & \vdots \\ \phi_1(\mathbf{x}_N) & \ldots & \phi_P(\mathbf{x}_N) \end{bmatrix}_{N\times P} \begin{bmatrix} \phi_1(\mathbf{x}_1) & \ldots & \phi_1(\mathbf{x}_N) \\ \vdots & \ddots & \vdots \\ \phi_P(\mathbf{x}_1) & \ldots & \phi_P(\mathbf{x}_N) \end{bmatrix}_{P\times N} \tag{8}$$

Now, remark that $\mathbf{\Phi}_{N\times P}$ corresponds to the sample matrix of the feature space. The objective of the kernel Principal Component Analysis, also known as kPCA, is to identify a linear projection which projects $\mathbf{\Phi}_{N\times P}$ onto uncorrelated

components, possibly of smaller dimensionality. One way to achieve that is to find a representation of $\mathbf{\Phi}_{N\times P}$ able to express each point $\phi_n \in \mathcal{F}$ as a linear combination of $Q \leq P$ orthogonal vectors of dimension P given as

$$\phi_n = \phi(\mathbf{x_n}) = \sum_{\mathbf{q=1}}^{\mathbf{Q}} \alpha_{\mathbf{n,q}} \mathbf{v_q} \tag{9}$$

such that for $\mathbf{v}_1 = [v_{q,1}, \ldots, v_{q,P}]$, we have

$$\mathbf{v}_q \mathbf{v}_k^\intercal = [v_{q,1}, \ldots, v_{q,P}] \begin{bmatrix} v_{k,1} \\ \cdots \\ v_{k,P} \end{bmatrix} = \sum_{p=1}^{P} v_{q,p} v_{v_k,p} = \begin{cases} 1 & \text{if } q = k \\ 0 & \text{otherwise} \end{cases}$$

and vectors $\boldsymbol{\alpha}_q = [\alpha_{1,q}, \ldots, \alpha_{N,q}]$ and $\boldsymbol{\alpha}_k = [\alpha_{1,k}, \ldots, \alpha_{N,k}]$ are uncorrelated that results in

$$\text{Cov}(\boldsymbol{\alpha}_q, \boldsymbol{\alpha}_k) = [\alpha_{1,q}, \ldots, \alpha_{N,q}] \begin{bmatrix} \alpha_{k,1} \\ \vdots \\ \alpha_{N,k} \end{bmatrix} = \sum_{n=1}^{N} \alpha_{n,q} \alpha_{n,k} = \begin{cases} \lambda_q & \text{if } q = k \\ 0 \text{ otherwise} \end{cases}$$

Let us define the matrix $\mathbf{A}_{n\times Q}$ and $\mathbf{V}_{Q\times P}$ such that

$$\mathbf{A}_{N\times Q} = [\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_Q] = \begin{bmatrix} \alpha_{1,1} & \ldots & \alpha_{1,Q} \\ \vdots & \ddots & \vdots \\ \alpha_{N,1} & \ldots & \alpha_{N,Q} \end{bmatrix}_{N\times Q} \tag{10}$$

and

$$\mathbf{V}_{N\times Q} = \begin{bmatrix} \mathbf{v_1} \\ \vdots \\ \mathbf{v_Q} \end{bmatrix} = \begin{bmatrix} v_{1,1} & \ldots & v_{1,P} \\ \vdots & \ddots & \vdots \\ v_{Q,1} & \ldots & v_{Q,P} \end{bmatrix}_{Q\times P} \tag{11}$$

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

$$\mathbf{A}^\intercal \mathbf{A} = \mathbf{\Lambda}_{Q\times Q} \quad \text{and} \quad \mathbf{V}\mathbf{V}^\intercal = \mathbb{I}_Q \tag{12}$$

and the representation that we are trying to find is defined as

$$\mathbf{\Phi}_{N\times P} = \mathbf{A}_{N\times Q} \mathbf{V}_{Q\times P} \tag{13}$$

If $Q < P$ we seek for an approximation of $\mathbf{\Phi}_{N\times K}$ in the euclidean norm that obtains the highest covariance.

2.2.2 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

$$\mathbf{C}_{P\times P} = \mathbf{\Phi}^\intercal_{P\times N} \mathbf{\Phi}_{N\times P} = \mathbf{V}^\intercal_{P\times Q} \mathbf{A}^\intercal_{Q\times N} \mathbf{A}_{N\times Q} \mathbf{V}_{Q\times P} = \mathbf{V}^\intercal_{P\times Q} \mathbf{\Lambda}_{Q\times Q} \mathbf{V}_{Q\times P}$$

and

$$\begin{aligned} \mathbf{C} &= \mathbf{V}^\intercal \mathbf{\Lambda} \mathbf{V} \\ \mathbf{V}\mathbf{C} &= \mathbf{V}\mathbf{V}^\intercal \mathbf{\Lambda} \mathbf{V} \\ \mathbf{V}\mathbf{C} &= \mathbf{\Lambda}\mathbf{V} \end{aligned}$$

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

$$\mathbf{\Phi}\mathbf{V}^\intercal = \mathbf{A}\mathbf{V}\mathbf{V}^\intercal \implies \mathbf{\Phi}\mathbf{V}^\intercal = \mathbf{A} \tag{14}$$

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

2.2.3 Unknown Feature Mapping ϕ

The feature mapping ϕ is usually unknown and, 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 known explicitly. One way to solve this problem is given by the employment of a kernel function $k(\cdot,\cdot)$, as given in Equation (7). Next, we show that the principal components $\mathbf{A}$ in the feature space and their corresponding variances stored on the diagonal of $\mathbf{\Lambda}$ can be obtained by using only the Gram Matrix $\mathbf{K}_{N\times N}$ defined through the inner product of the kernel. Since

$$\begin{aligned}\mathbf{C} &= \mathbf{V}^\intercal\mathbf{\Lambda}\mathbf{V}\\ \mathbf{V}\mathbf{\Phi}^\intercal\mathbf{\Phi} &= \underbrace{\mathbf{V}\mathbf{V}^\intercal}_{=\mathbb{I}_Q}\mathbf{\Lambda}\mathbf{V}\\ \mathbf{V}\mathbf{\Phi}^\intercal\underbrace{\mathbf{\Phi}\mathbf{\Phi}^\intercal}_{=\mathbf{K}_{N\times N}} &= \mathbf{\Lambda}\mathbf{V}\mathbf{\Phi}^\intercal\\ \mathbf{A}^\intercal\mathbf{K} &= \mathbf{\Lambda}\mathbf{A}^\intercal\end{aligned} \tag{15}$$

as $\mathbf{\Phi}\mathbf{V}^\intercal = \mathbf{A}$. Hence, 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

$$\mathbf{A}^\intercal\mathbf{A} = \mathbf{V}\mathbf{\Phi}^\intercal\mathbf{\Phi}\mathbf{V}^\intercal = \mathbf{V}\mathbf{C}\mathbf{V} = \mathbf{\Lambda} \tag{16}$$

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

$$\begin{aligned}\mathbf{\Lambda}^{-\frac{1}{2}}\mathbf{A}^\intercal\mathbf{K} &= \mathbf{\Lambda}^{-\frac{1}{2}}\mathbf{\Lambda}\mathbf{A}^\intercal\\ \mathbf{Z}^\intercal\mathbf{K} &= \mathbf{\Lambda}\mathbf{Z}^\intercal\end{aligned} \tag{17}$$

we have obtained orthonormal eigenvectors $\mathbf{Z}_{N\times Q} = \mathbf{A}\mathbf{\Lambda}^{-\frac{1}{2}} = \mathbf{\Phi}\mathbf{V}^\intercal\mathbf{\Lambda}^{-\frac{1}{2}}$ since $\mathbf{Z}^\intercal_{Q\times N}\mathbf{Z}_{N\times Q} = \mathbf{\Lambda}^{-\frac{1}{2}}\mathbf{\Lambda}\mathbf{\Lambda}^{-\frac{1}{2}} = \mathbb{I}_Q$. Therefore, by taking the eigendecomposition of the gram matrix $\mathbf{K}$ we obtain the matrices $\mathbf{Z}_{N\times Q}$ and $\mathbf{\Lambda}$ 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.

2.2.4 The Out-of-Sample Problem

As previously introduced, this problem comes into play once the feature mapping ϕ is either known or learnt and new sample points become the ones of interest in the analysis. In the first instance, hence if ϕ is known, then the principal components of the new sample $\phi(\mathbf{x}^\star)$ can be obtained as

$$\phi(\mathbf{x}^\star)\mathbf{V}^\intercal = \boldsymbol{\alpha}^\star \tag{18}$$

In the second case, instead, we need to define a new sample $\phi(\mathbf{x}^\star)$ by relying on the decomposition of $\mathbf{K}$ based on the initial sample $\phi(\mathbf{x}_1),\dots,\phi(\mathbf{x}_N)$. Let us recall that

$$\begin{aligned}\mathbf{\Phi} &= \mathbf{A}\mathbf{V}\\ \mathbf{A}^\intercal\mathbf{\Phi} &= \mathbf{A}^\intercal\mathbf{A}\mathbf{V}\\ \mathbf{A}^\intercal\mathbf{\Phi} &= \mathbf{\Lambda}\mathbf{V}\\ \mathbf{\Lambda}^{-1}\mathbf{A}^\intercal\mathbf{\Phi} &= \mathbf{V}_{Q\times P} = \begin{bmatrix}\mathbf{v}_1\\ \vdots\\ \mathbf{v}_Q\end{bmatrix}\end{aligned} \tag{19}$$

By setting $\mathbf{W}_{N\times Q} = \mathbf{A}_{N\times Q}\mathbf{\Lambda}^{-1}_{Q\times Q}$ we then achieve the formulation of the eigenvectors given the samples features $\mathbf{\Phi}$ and $\mathbf{W}$ that is

$$\mathbf{v}_q = \sum_{n=1}^{N} w_{n,q}\phi(\mathbf{x}_n) = \sum_{n=1}^{N}\frac{<\mathbf{v}_q,\phi(\mathbf{x}_n)>}{<\phi(\mathbf{x}_n),\phi(\mathbf{x}_n)>}\phi(\mathbf{x}_n) \tag{20}$$

and, on the other hand, since $\mathbf{\Phi} = \mathbf{A}\mathbf{V}$ and $\mathbf{\Phi}\mathbf{V}^\intercal = \mathbf{A}$ we have

$$\phi(\mathbf{x}_n) = \sum_{q=1}^{N} a_{n,q}\mathbf{v}_q = \sum_{n=1}^{N} <\mathbf{v}_q,\phi(\mathbf{x}_n)>\mathbf{v}_q \tag{21}$$

where the operator $< \cdot, \cdot >$ defines an inner product in $\mathcal{F}$ that in our case corresponds to the dot product, i.e. $< a, b >= \sum_{p=1}^{P} a_p b_p$. Following such definitions, we have the formulation of the projection based on the kernel functions as

$$k(\mathbf{x}_m, \mathbf{x}_n) \;=\; < \phi(\mathbf{x}_m), \phi(\mathbf{x}_n) > \;=\; \phi(\mathbf{x}_m)\phi(\mathbf{x}_n)^\intercal \tag{22}$$

since

$$a_{m,q} = \phi(\mathbf{x}_m)\mathbf{v}_q^\intercal = \phi(\mathbf{x}_m)\sum_{n=1}^{N} w_{n,q}\phi(\mathbf{x}_n)^\intercal = \sum_{n=1}^{N} w_{n,q}\phi(\mathbf{x}_m)\phi(\mathbf{x}_n)^\intercal = \sum_{n=1}^{N} w_{n,q}k(\mathbf{x}_m, \mathbf{x}_n) \tag{23}$$

Therefore, given a new sample data point in the feature space $\phi(\mathbf{x}^\star)$ for $\mathbf{x}^\star \in \mathcal{Y}$, we have its projection specified only by the kernel function and the eigendecomposition of $\mathbf{K}$, that is

$$a_q^\star \;=\; \phi(\mathbf{x}^\star)\mathbf{v}_q^\intercal = \sum_{n=1}^{N} w_{n,q}k(\mathbf{x}^\star, \mathbf{x}_n) \tag{24}$$

2.2.5 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 one. We review this concept since we will be using such a method for the hyperparameter learning procedure presented in subsection 3.2.3.

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 or, even if this was not the case, the inversion could lead to a pre-image $\tilde{\mathbf{x}}$ which is not unique. This is indeed an ill-posed problem where one seeks instead for 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

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

Several solutions have been proposed in the literature. The reader might refer to [34] to review some of the explored approaches, which we partly summarised in Appendix A. We employ the one proposed in [35]. To introduce it, remark first that the function $\phi(\cdot)$ is defined on a vector space. Thus, 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

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

where the operator P_k highlights that such a projection is induced through the kernel $k(\cdot, \cdot)$. Note that if Q is big enough then $\phi(\mathbf{x}) \approx P_k\phi(\mathbf{x})$. 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

$$\hat{\mathbf{x}} = \text{argmin}_{\mathbf{x}\in\mathcal{X}} \|P_k\phi(\mathbf{x}) - \phi(\hat{\mathbf{x}})\|_{\mathcal{F}}^2 \tag{25}$$

In practice, the pre-image problem searches for a map Γ with the property that $\Gamma(\phi(\mathbf{x}_i)) = \mathbf{x}_i$ for $i = 1, \ldots, N$. The second remark required to apply this method is that the pre-image map Γ can 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

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

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

$$\Gamma_j(P_k\phi(\mathbf{x})) = \sum_{i'=1}^{N} \beta_{i'}^j k'(P_k\phi(\mathbf{x}), P_k\phi(\mathbf{x}_{i'})) \tag{26}$$

and $k'(\cdot, \cdot)$ is a new kernel function which differs 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 $k'(\cdot, \cdot)$ 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 [36]) which consists of minimising the following function in its dual form below presented:

$$\hat{\mathbf{\Gamma}} = \text{argmin}_{\mathbf{\Gamma}} \sum_{i=1}^{N} l\left(\mathbf{x}_i - \mathbf{\Gamma}(P_k\phi(\mathbf{x}_i))\right) + \lambda R(\mathbf{\Gamma}) \tag{27}$$

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

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

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

$$\Big(\mathbf{x}_i - \mathbf{\Gamma}(P_k\phi(\mathbf{x}_i)\Big) = \|\mathbf{x}_i - \mathbf{\Gamma}(P_k\phi(\mathbf{x}_i))\|^2$$

and the next regularisation form

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

the criterion exploiting kernel ridge regression can be reformulated in its dual form (see [36] for derivation and details) and the solution is given as

$$\begin{aligned} \hat{\mathbf{B}}_{N\times D} &= \text{argmin}_{\mathbf{B}}\ \text{tr}\left([\mathbf{X} - \mathbf{K}'\mathbf{B}][\mathbf{Y} - \mathbf{K}'\mathbf{B}]^\intercal\right) + \lambda \text{tr}(\mathbf{B}\mathbf{B}^\intercal) \\ &= \left(\mathbf{K}'^\intercal\mathbf{K}' + \lambda\, \mathbf{I}_N\right)^{-1} \mathbf{K}'^\intercal\mathbf{X} \end{aligned} \tag{28}$$

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

$$\mathbf{K}'_{i,j} = k'(P_k\phi(\mathbf{x}_i), P_k\phi(\mathbf{x}_j)) = \phi'(P_k\phi(\mathbf{x}_i))\ \phi'(P_k\phi(\mathbf{x}_j))^\intercal \tag{29}$$

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

$$\mathbf{K}'_{i,j} = k'(a_{i,q}, a_{j,q}) = \phi'(a_{i,q})\ \phi'(a_{j,q})^\intercal \tag{30}$$

and hence we have

$$bfK' = \begin{bmatrix} \phi_1^{'}(a_1) & \ldots & \phi_Q^{'}(a_1) \\ \vdots & \ddots & \vdots \\ \phi_1^{'}(a_N) & \ldots & \phi_Q^{'}(a_N) \end{bmatrix}_{N\times Q} \begin{bmatrix} \phi_1^{'}(a_1) & \ldots & \phi_1^{'}(a_N) \\ \vdots & \ddots & \vdots \\ \phi_Q^{'}(a_1) & \ldots & \phi_Q^{'}(a_N) \end{bmatrix}_{Q\times N}$$

Define now the following vector

$$\mathbf{k}_{\mathbf{x}}^{'} = [k'(P_k\phi(\mathbf{y}), P_k\phi(\mathbf{x}_1)), \ldots, k'(P_k\phi(\mathbf{x}), P_k\phi(\mathbf{x}_N))]_{1\times N} \tag{31}$$

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

$$\hat{\mathbf{x}} = \mathbf{\Gamma}(P_k\phi(\mathbf{x})) = (\mathbf{k}_{\mathbf{x}}^{'}\hat{\mathbf{B}})_{\ 1\times D} \tag{32}$$

2.3 CCA for Linear Cross-Correlation between Two Data sets

In this section, we review CCA and how it is working, how it is derived along with its interpretation. Readers might refer to [37, 38] for further details and discussions about the interpretation (and references within). 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. Indeed, most of dependence methods are special cases of CCA, as multiple regression with only one response variable, two-group discriminant with one dummy variable as response, multi-group discriminats with several dummy variables as responses, ANOVA with only one response and dummy variables as indep and MANOVA with several

responses and dummy variables as indep. It can subsume these tests as special cases within the General Linear Model (GLM). Understanding these advantages can aid in selecting appropriate statistical techniques and enhance conceptual understanding throughout the GLM framework.

The main appeal of this multivariate method comes from the interest in finding an association between two data sets. If the simple correlation matrix is considered, one will obtain the sample correlations between all pairs of variables without having information for assessing the within-set associations and the between-set associations (aka 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 insight relationships between the two data sets that are hidden and affected by the within-set. Therefore, CCA seeks linear combinations of the first data set and linear combinations of the second data set, which are more highly correlated. Since the objective 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. Such a sought result can be obtained by imposing each pair of linear combinations to be 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

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

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

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

where $\mathbf{\Sigma}_{XX}$ is the $d' \times d'$ sample correlation matrix of the first sets of variables, $\mathbf{\Sigma}_{YY}$ is the $d \times d$ sample correlation matrix of the second sets of variables, $\mathbf{\Sigma}_{XY}$ is the $d' \times d$ sample matrix of correlations between the variables of $\mathbf{X}$ and $\mathbf{Y}$. $\mathbf{\Sigma}_{YX}$ corresponds to the transpose of $\mathbf{\Sigma}_{XY}$. 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 member of U_i is paired with a member 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 i-th pair (U_i, V_i) corresponds to the canonical variate. CCA seeks linear combinations that maximise the correlations between the members of each pair of canonical variates. The correlation between U_i and V_j is obtained as usual by standard liner correlation measures

$$Corr\left(U_i, V_j\right) = \frac{Cov\left(U_i, V_j\right)}{\sqrt{Var\left(U_i\right) Var\left(V_j\right)}}$$

The canonical correlation is the specific type of correlation for the pair of i-th canonical variate given by the correlation between U_i and V_i that will be denoted by ρ_i^*

$$\rho_i^* = \frac{Cov\left(U_i, V_i\right)}{\sqrt{Var\left(U_i\right) Var\left(V_i\right)}}$$

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

$$\boldsymbol{a}^{(j)}, \boldsymbol{b}^{(j)} = \underset{\boldsymbol{a},\boldsymbol{b}}{\operatorname{argmax}} \operatorname{Corr}\left(\boldsymbol{a}^T \boldsymbol{X}_i, \boldsymbol{b}^T \boldsymbol{Y}_i\right)$$

subject to constraint set

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

Given linear algebra solutions, in practice, the vectors $\mathbf{a}_i$ and $\mathbf{b}_i$, corresponding to the coefficient of the linear combinations defining the pairs (U_i, V_i) respectively, can be obtained as follows. The coefficients $\mathbf{a}_i$ are the eigenvectors of the matrix

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

and $\mathbf{b}_i$ are the eigenvectors of the matrix

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

The canonical correlation will correspond to the square roots of the non-zero eigenvalues of the above matrices. Furthermore, to assess the statistical significance of the canonical correlations the null hypothesis is given as

$$\begin{aligned} &H_0 : \rho_1^* = \rho_2^* = \cdots = \rho_{d'}^* = 0 \\ &H_1 : \text{At least one canonical correlation significantly differs from zero} \end{aligned}$$

For testing the above, for testing the above hypotheses, the most widely used test statistic is Wilks' lambda, given by $\Lambda = \prod_{i=1}^{d'}(1 - \rho_i^2)$. Bartlett showed that under the null hypothesis a particular function of Λ would be distributed approximately as a chi-squared variate [39]. Therefore the statistical significance of Wilk's Λ requires the calculation of the following statistic:

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

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 a sequence of procedures that test whether the second-largest canonical correlation significantly differs from zero, then the third largest, etc. 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 [40]. This employs a likelihood ratio test approach in the latent space [**?**]. Starting with $i = 0$, the null hypothesis tests

$$\begin{aligned} &H_0 : d = i \\ &H_1 : d > i \end{aligned}$$

If H_0 is rejected, i is incremented and a new test is conducted. This proceeds until H_0 is not rejected or i reaches $M = min(d, d')$. For a given number of canonical variates, m, the Wilk's test statistics is given (more formally than above) by

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

Based on Rao's F-approximation the test follows $F_{\text{df}_1,\text{df}_2}$ given as

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

where $\nu = \sqrt{(\text{df}_1^2 - 4)/((d' - m)^2 + (d - m)^2 - 5)}$, $\text{df}_1 = (d' - m)(d - m)$ and $\text{df}_2 = (N - 1.5 - (d' + d)/2)\nu - \text{df}_1/2 + 1$ (where N is the numner of samples). Such a test will be the one employed to evaluate the CCA models within this work.

Given such a framework, then 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 $\rho_1^\star, \rho_2^\star, \ldots, \rho_{d'}^\star$ (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. The canonical variate are therefore somewhat artificial and difficult to interpret. The

standard way to observe results of CCA is by looking at the coefficients to understand which variables provide the highest loadings on the (U_i, V_i) and so describe the associations between sets of variables. Remark that associations that does not imply any causal relationship across the two data sets.

In this regard, several quantities must be considered when evaluating the CCA model and providing its interpretation. 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. Another quantity often employed 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 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 modeling, 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. As introduced, 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 [41, 42]. Lastly, the *Squared canonical structure coefficients* are the square of the structure coefficients. This statistic is analogous to any other r2-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.

CCA Model Assessment

Quantity	Notation	Interpretation	Relation with Other Models
Canonical Correlation Coeff.	ρ_i^*	Association between $\mathbf{X}$ and $\mathbf{Y}$	Similar to the multiple R value in regression
Squared Canonical Correlation	$\rho_i^{*^2}$	Proportion of shared variance between the synthetic variables in each canonical function	Analogous to the R^2 effect in multiple regression
Redundancy Index	$(\sum_{j=1}^{d'} Corr^2(Y_j, V_i)/d')\rho_j^{*^2}$	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 $\mathbf{X}$ and $\mathbf{Y}$	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)	$Corr(X_i, U_j)$, $Corr(Y_i, V_j)$	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.	$Corr^2(X_i, U_j)$ $Corr^2(Y_i, V_j)$	Proportion of variance an observed variable linearly shares with the synthetic variable	Analogous to any other R^2-type effect size

Table 1: Quantities required for assessing the CCA model.

Once the quantities in Table 1 are computed, the general approach in the analysis follows these steps: (1) the canonical functions are estimated, and the magnitudes of the canonical correlation coefficients are quantified along with what is usually referred to as the redundancy index. (2) The second part required to proceed in the analysis consists of assessing the model as a whole through several statistical tests. (3) If these results are statistically significant, one can then move to the interpretation of the relative importance of each of the canonical functions by using standardised canonical coefficients (i.e. canonical weights) and canonical loadings (i.e. structure correlations). (4) The last step considered might be using orthogonal rotation to facilitate the interpretation of canonical functions, canonical loadings, and standardised canonical coefficients. The model can then be validated. The procedure followed in the analysis of the results followed such a reasoning.

At this stage there is a relevant point that must be taken into account. For the CCA to work, the underlying data sets must be linearly associated and, furthermore, the assumption of multi-normality must be respected by $\mathbf{X}$ and $\mathbf{Y}$, i.e. the variables in each set are normally distributed. Reasons behind that are firstly the validity of the statistical tests performed when the assessment of the overall model is made. Secondly, the interpretability of the results, since, when the variables are normally distributed, it is easier to understand the relationship between the canonical variates and the original variables. Non-normality can complicate the interpretation of the canonical weights and make it harder to understand the underlying relationships. Furthermore, multi-normality assumptions can improve the accuracy of parameter estimation in CCA. Under normality, maximum likelihood estimation is often used to estimate the canonical correlation coefficients and canonical weights. Violations of the multi-normality assumption may lead to biased estimates or unreliable results. Lastly, a better statistical efficiency might be achieved with multi-normality, meaning that it provides the most precise estimates and optimal power for hypothesis testing. Departures from normality can reduce statistical efficiency, leading to less precise estimates and decreased power.

In practice, when it comes to complex real data scenarios, this kind of assumptions will rarely be encountered. Highly non-linear and non-stationary data sets are analysed, and, these two properties affect the multi-normality assumption leading to distributions of which tend to be heavy-tail distributions or lacking of symmetry or with prominent extreme values and leading the results towards unreliable estimates and statistical testings. Therefore, some robust alternatives to CCA has been proposed in the literature [43] that relax the assumption of multi-normality. These methods, such as robust canonical correlation analysis [43], are designed to handle non-normal or non-linear data. The idea of this work is to propose a method following this class of robust solutions and facilitate the job of CCA by capturing the non-stationary nature of the underlying data for better interpretation of the obtained cross-correlations. This is achieved by firstly applying the kPCA and extract a set of kPCs to then capture the cross-correlation of these basis functions. To better understand this point, we derive a toy example in subsection 3.3, showing the mathematical steps of such an idea.

3 Non-Stationary Leading Factors Extraction for Robust Cross-Correlation Detection

In this section, we present the methodology of the paper. It proposes a novel solution to detect cross-correlation between multiple multivariate data sets in an automated fashion, which will be robust to the presence of non-stationary and non-linear signals or time series as well as the presence of different structured data, i.e. numerical or categorical or different timestamps observation frequencies and spatial locations. The procedure exploits kPCA for extracting non-stationary and non-linear basis components, which, relying on kernel functions, gives reliable information about the variability of the original complex data structure. Each non-linear basis indicates the amount of captured variation provided by its associated eigenvalue. Therefore, the advantages of the kPCs are (1) handling non-linearity and non-stationarity of the data and (2) automatically providing a threshold, i.e. the eigenvalues, for reconstructing the data without the need for spectral truncation as standard singular value decomposition or alternative methods ([44]).

Multiple 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 that county. This approach results in a multivariate data set available for all Californian counties. The primary goal is to comprehensively characterise the variations present in these multi-multivariate data sets simultaneously, treating the fluctuations of 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.

The impact of the green bonds given is intended to be detected around highly polluted and populated areas. Therefore, as highlighted in subsection 1.1, we split the originals data sets as $\mathbf{X}^1_{N_1\times D_1} = [\mathbf{x}^{1,1}_{T_1\times D_1}, \mathbf{x}^{1,2}_{T_1\times D_1}, \dots, \mathbf{x}^{1,S}_{T_1\times D_1}]$, $\mathbf{X}^1_{N_1\times D_1} = [\mathbf{x}^{1,1}_{T_1\times D_1}, \mathbf{x}^{1,2}_{T_1\times D_1}, \dots, \mathbf{x}^{1,S}_{T_1\times D_1}]$ and $\mathbf{X}^3_{N_3\times D_3} = [\mathbf{x}^{3,1}_{n_1\times D_3}, \mathbf{x}^{3,2}_{n_2\times D_3}, \dots, \mathbf{x}^{3,S}_{n_J\times D_3}]$ where n_j for $S = 1, \dots, S$. $\mathbf{X}^1_{N_1\times D_1} = [\mathbf{x}^{1,1}_{T_1\times D_1}, \mathbf{x}^{1,2}_{T_1\times D_1}, \dots, \mathbf{x}^{1,S}_{T_1\times D_1}]$, where T_1 is the number of timestamps per county and the upper indices indicate the first data set (pollution) and the number of counties S, respectively. This reasoning is applied to the second data set as well, i.e. the climate one. In the case of the financial data, we will have $\mathbf{X}^3_{N_3\times D_3} = [\mathbf{x}^{3,1}_{n_1\times D_3}, \mathbf{x}^{3,2}_{n_2\times D_3}, \dots, \mathbf{x}^{3,J}_{n_J\times D_3}]$ where n_s for $s = 1, \dots, S$ is the number of issued and active green bonds per county. The kPCs will be then evaluated at new feature points, i.e. on a new mesh, common across the counties, as per Equation 24. By exploiting the out-of-sample problem presented in subsection 2.2.4, we will obtain a new set of kPCs which are comparable since based on the same grid. The final step consists of first employing the CCA on the kPCs and observing the captured cross-correlations. Fig. 1 summarises this implemented methodology.

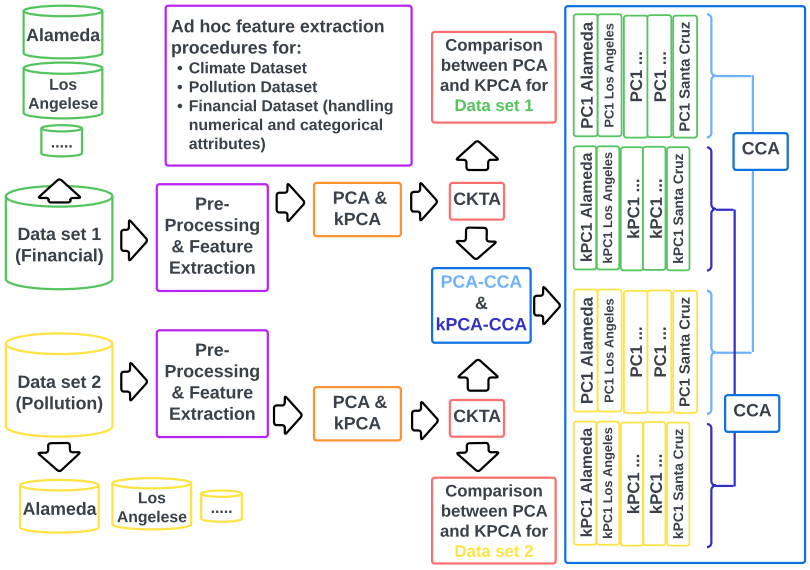


Figure 1: Figure presenting the steps of the proposed method. Firstly, the kPCA-CCA and its benchmark PCA-CCA are run to observe municipal green bonds impact on pollution and climate. Therefore, the analysis will be performed by running kPCA-CCA (and PCA-CCA) on pollution vs financial kPCs (PCs) and climate vs financial kPCs (PCs). The Figure is set on the first case, but we perform an equivalent analysis for climate. Hence, we will describe the figure on the pair pollution/financial and consider similar steps for the case climate/financial. Given the two data set, pollution and financial data set, we split them by county and perform ad hoc cleaning and pre-processing procedures.Afterwards, a specific set of features have been extracted depending on the data. In the case of pollution, this has foreseen engineering ad hoc features summarising information across several time series collected at different monitors. In the case of financial, numerical and categorical variables have been handled. After, both the linear PCA and the non-linear kPCA have been applied and, for each decomposition method, the first three bases (PC1, PC2, PC3, kPC1, kPC2, kPC3) have been retained. Once obtained, the next step consisted of comparing which bases better captured the variability of the given data set through 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.

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 Subsection 2.2.5. 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. Figure 1 shows the steps for the proposed methodology applied to two data sets with the added aspect of the Jaccard kernel for the financial one.

The toy example given in subsection 3.3 will show the mathematical steps applied to obtain the CCA on the original data, on the PCs and on the kPCs respectively, and what are the difference in applying such a procedure. In practice, in the case of PCA-CCA, we will isolate the most relevant content in terms of variation of the underlying data set by spectral ordering and then the content can be used. The CCA will then recover the dominant component of the marginal parts. In the case of the kPCA,the interpretation is that we have capture the averaging in time and space. The final goal is to capture the averaged time and spatial instantaneous correlation of the leading spectral components or leading eigen-functions through time.

The remaining of this section is therefore organised as follows: firstly we review the procedure for the definition of a new mesh required for the evaluation of the kPCs on a common grid, making it comparable. Second, we review kernel functions and kernel learning procedure with a description of the radial basis function and the Jaccard kernels and, afterwards, the hyperparmater learning procedure is presented. Lastly, a statistical toy example showing the interpretation of PCA-CCA and kPCA-CCA is provided.

3.1 The Definition of a New Mesh

Once obtained, the optimal KPCs of each dataset $\{\{\mathbf{x}^{i,s}_{T_i\times D_i}\}^S_{s=1}\}^3_{i=1}$ are extracted at points belonging to different input spaces, formally $\mathcal{X}^s_i$, for $i = 1, 2, 3$ and $s = 1, \dots, S$. The following part of the implemented methodology will be searching for associations at different modes of variation (provided by the KPCs) which are, however, evaluated at different input space sample points and, therefore, cannot be compared. Figure 2 shows the point of this argument and justifies the need for a new common mesh at which each kernel principal function is evaluated. In such a way, we will define a new set of KPCs, by exploiting the out-of-sample problem presented in 2.2.4, which will then be comparable and used as input to the the CCA.

For each ith dataset, we define a new mesh of points $\mathbf{x}^{\star,i}$ common across the J counties. Consider the pollution dataset divided by county $\{\mathbf{x}^{1,s}_{T_1\times D_1}\}^S_{s=1}$. To define a new mesh, we applied the following procedure. Compute the maximum and the minimum for each column

$$\begin{aligned}\max_{D_1}\{\mathbf{x}^{1,s}_{T_1\times D_1}\}^S_{s=1} &= \{[M^s_1, M^s_2, \dots, M^s_{D_1}]\}^S_{s=1} = \{\mathbf{M}^s_{1\times D_1}\}^S_{s=1}\\ \min_{D_1}\{\mathbf{x}^{1,s}_{T_1\times D_1}\}^S_{s=1} &= \{[m^s_1, m^s_2, \dots, m^s_{D_1}]\}^S_{s=1} = \{\mathbf{m}^s_{1\times D_1}\}^S_{s=1}\end{aligned} \tag{33}$$

Afterwards, we take the minimum of the maxima and the maximum of the minima as follows

$$\begin{aligned}\min_j\{\mathbf{M}^s_{1\times D_1}\}^S_{s=1} &= [M_1, M_2, \dots, M_{D_1}]\\ \max_j\{\mathbf{m}^s_{1\times D_1}\}^S_{s=1} &= m_1 = [m_1, m_2, \dots, m_{D_1}]\end{aligned} \tag{34}$$

We then generate a uniform grid of $L = 100$ new input points in the intervals $I_1 = [m_1, M_1], I_2 = [m_2, M_2], \dots, I_{D_1} = [m_{D_1}, M_{D_1}]$. Afterwards. we can evaluate the formula for $a^\star_q$ for this dataset and each county by using a common input matrix defined as $\mathbf{x}^{\star,1}_{L\times D_1}$, where $L = 100$. Note that we retain three new coordinates, i.e. one for the first three eigenvectors of the Kpca $q = 1, 2, 3$ which we denote as

$$a^{\star,1,s}_{l,q} = \phi^{1,s}\left(\mathbf{x}^{\star,1}_{l\times D_1}\right)\mathbf{v}^{\intercal,1,s}_q = \sum_{n=1}^{n_1} w^{1,s}_{n,q}\ k^{1,s}(\mathbf{x}^{\star,1}_{l\times D_1}\ ,\ \mathbf{x}^{1,s}_n) \tag{35}$$

where $s = 1, \dots, S$ is the index related to the county, $l = 1, \dots, L$ are the number of points of the new mesh and the index 1 indicates the first dataset. Therefore, we denote each of the new coordinate vectors as

$$\mathbf{a}^{\star,1,s}_q = [a^{\star,1,s}_{1,q}, a^{\star,1,s}_{2,q}, \dots, a^{\star,1,s}_{L,q}]_{1\times L} \tag{36}$$

for $q = 1, 2, 3$. We will use each of these vectors to construct the graphs. The details will be presented within the sections below. We apply the same procedure to the other dataset and therefore obtain

$$\begin{aligned}\mathbf{a}^{\star,2,s}_q &= [a^{\star,2,s}_{1,q}, a^{\star,2,s}_{2,q}, \dots, a^{\star,2,s}_{L,q}]_{1\times L}\\ \mathbf{a}^{\star,3,s}_q &= [a^{\star,3,s}_{1,q}, a^{\star,3,s}_{2,q}, \dots, a^{\star,3,s}_{L,q}]_{1\times L}\end{aligned} \tag{37}$$

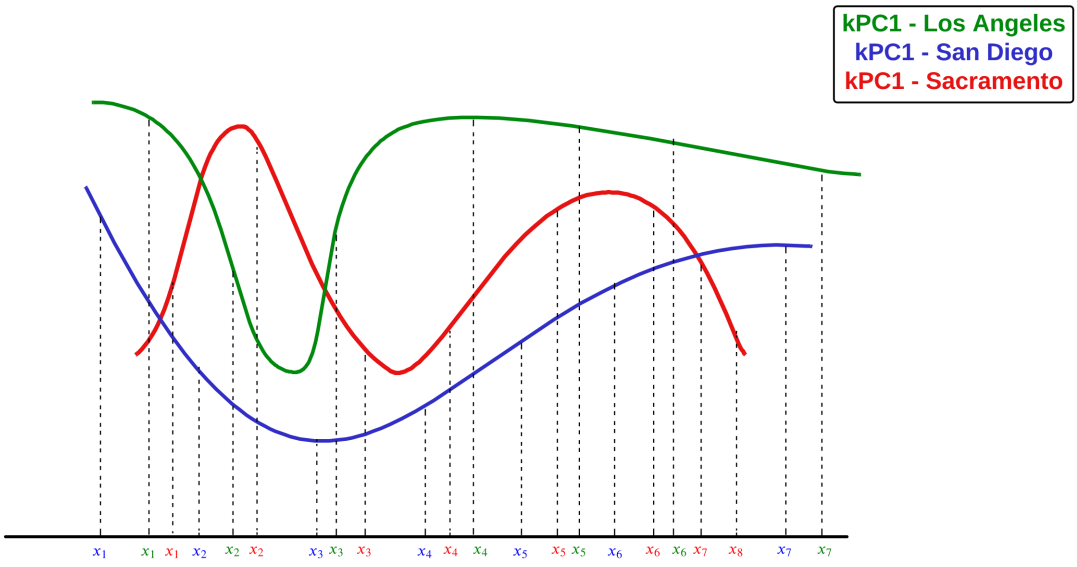


Figure 2: Examples presenting the need for the new mesh. Consider kPC1, hence the first kernel principal component extracted in the counties of Los Angeles, San Diego and Sacramento as given. We collected data from different monitors and averaged them across different grids as shown in the above picture, i.e. we have one grid per county provided in different colors. Therefore, when comparing the kPCs, a unique mesh where these bases are evaluated should be considered.

3.2 Kernel Functions and Learning

So far we have discussed and introduced the Kpca in general for each of the input matrices $\{\mathbf{X}^i_{N_i\times D_i}\}_{i=1}^3$ simply denoted as $\mathbf{X}$. At this stage it is important to observe that we apply this technique by splitting the original dataset by counties in order to observe and define association between green bonds and pollution or green bonds and climate according in areas which are highly populated and so strongly affecting life conditions. Therefore, we learn learn a feature mapping $\phi^{i,s}$ that characterise a specific county s, with $s = 1, \ldots, S$ for each of the datasets i, with $i = 1, 2, 3$. This means that we apply the Kpca to each of the input matrices divided by county, i.e. for the pollution dataset, for example, this is $\mathbf{X}^1_{N_1\times D_1} = [\mathbf{x}^{1,1}_{T_1\times D_1}, \mathbf{x}^{1,2}_{T_1\times D_1}, \ldots, \mathbf{x}^{1,S}_{T_1\times D_1}] = \{\mathbf{x}^{1,s}_{T_1\times D_1}\}_{s=1}^S$; and, equivalently for the remaining dataset, we have $\{\mathbf{x}^{2,s}_{T_2\times D_2}\}_{s=1}^S$ and $\{\mathbf{x}^{3,s}_{T_3\times D_3}\}_{s=1}^S$. As a result, the kernel trick is used to learn a set of feature mappings defined as $\{\phi^{i,s}\}_{s=1}^S$, which have different input sets of sample points $\{\mathbf{x}^{i,s}_{T_i\times D_i}\}_{s=1}^S$, for $i = 1, \ldots, 3$ and $s = 1, \ldots, S$.

3.2.1 The Radial Basis Function Kernel for Numerical Variables

Hence, the hyperparameter learning procedure will be applied to $\{\{\mathbf{x}^{i,s}_{T_i\times D_i}\}_{s=1}^S\}_{i=1}^3$, using the radial basis kernel function, whose formulation for a general $\mathbf{x}$ is given as

$$k_{\text{rbf}}(\mathbf{x}, \mathbf{x}') = \exp\left(\frac{-|\mathbf{x} - \mathbf{x}'|^2}{2\sigma^2}\right) = \exp(-\gamma\,|\mathbf{x} - \mathbf{x}'|^2) \tag{38}$$

where $\gamma = \frac{1}{2\sigma^2}$.

3.2.2 The Jaccard Kernel for Categorical Variables

The financial dataset referring to municipal green bonds contains multiple data sources. Indeed, certain variables are numerical, while many others are categorical instead. Examples of numerical variables are coupon, maturity size, spread, spread duration, etc., whereas categorical variables are the type of municipal bond issued, the industry of the issued bond, etc. The complete description of this data is provided in subsection 4.3. The steps for the extraction of the KPCs have been described above. As highlighted, the optimal KPCs are computed according to the hyperparameters learning procedure given in subsection 3.2.3, which takes into account the pre-image problem. In our case, the

employed kernel has been a radial basis function kernel of Equation (38). This function is applicable if and only if the underlying data are numerical. Hence, for several financial variables, this choice cannot be considered. To proceed further, we then took into account two main considerations.

The first one foresees the fact that the sum of two positive definite kernels is a positive definite kernel. Hence, if $k_1(\mathbf{x}, \mathbf{x}')$ and $k_2(\mathbf{x}, \mathbf{x}')$ are two different kernel functions then

$$k(\mathbf{x}, \mathbf{x}') = k_1(\mathbf{x}, \mathbf{x}') \ + \ k_2(\mathbf{x}, \mathbf{x}') \quad \forall \ \mathbf{x}, \mathbf{x}'$$

$k(\mathbf{x}, \mathbf{x}')$ is also a kernel function. Therefore, if we split the financial variables amongst the numerical and the categorical ones, we can end up with a unique kernel for such a dataset by multiplying the Gram Matrix for the numerical variables and the Gram Matrix for the categorical. Since the kernel for the numerical variables is indeed the radial basis function, this subsection presents the kernel function required for the categorical variables. This is achieved by considering the Jaccard index ([45], [46]), also known as Jaccard coefficient and defined as

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|} \tag{39}$$

where A and B are any two sets to be compared. It is henceforth understood that $J(A, B) = 0$ in the case of $|A \cap B| = 0$ and $|A \cup B| = 0$. Note, furthermore that $0 \leq J(A, B) \leq 1$. Though not frequently specified, the universe set of A and B can be conveniently taken as being equal to $\Omega = A \cup B$. Such index measure the similarity between finite sample sets. The Jaccard distance can be easily derived from the Jaccard index and given by

$$d_J(A, B) = 1 - J(A, B) = \frac{|A \cup B| - |A \cap B|}{|A \cup B|} \tag{40}$$

This approach can be immediately extended to any other similarity index bound between 0 and 1. Moreover, it is possible to observe that this measures dissimilarity between sample sets, complementary to the Jaccard coefficient given in Equation (39).

Therefore, the final kernel for the financial data will be given as

$$k(\mathbf{x}, \mathbf{x}') = k_{\text{rbf}}(\mathbf{x}, \mathbf{x}') \ + \ k_{\text{jaccard}}(\mathbf{x}, \mathbf{x}') \tag{41}$$

3.2.3 Hyperparameter Learning Procedure

In this section, we review the procedure for the Kpca hyperparameter learning. This is required since each kernel function carries a set of hyperparameters which are unknown and, therefore, have to be learnt from the available dataset. Once the optimal hyperparameters are identified, the optimal KPCs can be extracted and used in the KPCA-CCA methodology. There are several methods for hyperparameters learning. We chose a grid-search given its simplicity and straightforward interpretation. Before introducing the procedure, we need to make some remarks.

The learning involves a grid-search made to select the optimal hyperparameter γ for each of the datasets $\{\{\mathbf{x}^{i,j}_{n_i \times D_i}\}_{j=1}^{47}\}_{i=1}^{3}$. The procedure exploits the pre-image method above introduced. Note that given the formulation of the optimisazion problem, we also conducted a grid search over the parameter λ controlling the regularisation term. We set the two grids as $\gamma = (0.01, 0.1, 0.5, 1, 5, 10, 30, 50)$ and $\lambda = (0.01, 0.1, 1, 10, 100)$. The idea is mapping a single dataset into the feature space and then computing its pre-image for each of the hyperparameters. Afterwards, the Euclidean distance of the original dataset and the obtained pre-image is calculated. Let us select one single dataset $\mathbf{x}^{i,s}_{n_i \times D_i}$ and denote its entries:

$$\mathbf{x}^{i,s}_{T_i \times D_i} = \begin{bmatrix} x^{i,s}_{1,1} & \dots & x^{i,s}_{1,D_i} \\ \vdots & \ddots & \vdots \\ x^{i,s}_{T_i,1} & \dots & x^{i,s}_{T_i,D_i} \end{bmatrix}_{T_i \times D_i}$$

The Euclidean distance is calculated rowise on the above matrix and its correspondent pre-image $\hat{\mathbf{x}}^{i,s}_{n_i \times D_i}$. Let us slightly modify our notation as $\mathbf{x}^{i,s}_{T_i,D_i}$ so to make use of the lower indices. The Euclidean distance is then given as follows

$$\mathrm{d}(\mathbf{x}^{i,s}_{T_i \times D_i}, \hat{\mathbf{x}}^{i,s}_{T_i,D_i}) = \frac{1}{D_i} \sum_{D_i=1}^{D_i} \left(\mathbf{x}^{i,s}_{T_i,D_i} - \hat{\mathbf{x}}^{i,s}_{T_i,D_i} \right)^2 \tag{42}$$

Therefore, for each $\mathbf{x}^{i,s}_{T_i \times D_i}$ we compute 8×5 Kpca, one for each combination of grid values given by the hyperparameters, and select the optimal $\hat{\mathbf{x}}^{i,s}_{T_i \times D_i}$ which is the one minimizing the above criterion. Once the optimal hyperpameter

set is identified, then the principal components for that set are extracted. We summarise the procedure. For each dataset divided by county $\mathbf{x}^{i,s}_{T_i \times D_i}$, each values of λ and γ, apply the following

Algorithm 1: Hyperparameter Learning Algorithm

Input: $\mathbf{x}^{i,s}_{T_i \times D_i}, \lambda, \gamma$

for *i,s,*λ*,* γ **do**

1. Evaluate the KPCs given within the matrix $\mathbf{A}^{i,s}$
2. Compute the pre-image $\hat{\mathbf{x}}^{i,s}_{T_i \times D_i}$;
3. Compute the Mean Square Error of the pre-image and the original dataset as given in Equation (42).
4. Extract the optimal KPCs according to the minimum Euclidean distance

3.3 Toy Example

The final goal of this example is to reproduce a simple case study where we provide a straightforward interpretation of why we combine the CCA along with the kPCA. However, before getting to the kernel version, we firstly consider a linear example, i.e. an example where we employ the PCA. We remark that the purpose of our methodology is capturing cross-correlation amongst the given datasets, i.e. green bonds data versus climate data or green bonds data versus pollution data. The relevant aspects of this example go as follows: firstly, standard CCA is derived for a toy example considering two data sets $\mathbf{X}$ and $\mathbf{Y}$. Once the canonical variates are defined, we then apply PCA to $\mathbf{X}$ and $\mathbf{Y}$ and develop the solution of PCA-CCA. Lastly, we employ kPCA for extracting the kPCs of $\mathbf{X}$ and $\mathbf{Y}$ and then construct the kPCA-CCA solution.

The example is now presented. Consider the two data sets $\mathbf{X}_{N \times d'}$ and $\mathbf{Y}_{N \times d}$. The first case that we want to observe is applying the CCA directly to the raw data. The CCA seeks two vectors $\mathbf{a} \in \mathbb{R}^{d'}$ and $\mathbf{b} \in \mathbb{R}^{d}$ such that the linear combinations $\mathbf{X}_i \mathbf{a}^\intercal$ and $\mathbf{Y}_i \mathbf{b}^\intercal$ have maximum correlation given as

$$\rho(\mathbf{a}, \mathbf{b}) = \rho(\mathbf{X}\mathbf{a}^\intercal \, \mathbf{Y}\mathbf{b}^\intercal)$$

Note that in the CCA subsection 2.3 the correlation is expressed as $\rho(\mathbf{a}^\intercal \mathbf{X}\mathbf{b}^\intercal \mathbf{Y})$, while in the above the vectors are multiplied with the matrix rather than the opposite. This inversion ca be performed without loss of generality. Suppose now one has

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

with

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

Hence, one can write

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

Note that the property of scale invariance applies as follows

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

The CCA problem can then be solved as

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

where the constraints equal to one provide the fact that we look for the maximum correlation. Define

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

If one now applies SVD then they will obtain

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

where

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

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. For $i \in 1, \ldots, k$ we have

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

which corresponds to the canonical correlation vectors. From these, we can then obtain

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

which are the canonical correlation variables in 2D. 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 $\mathbf{Y}_{N \times d}$ and is obtained by linear correlation coefficients. 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 coeffcients 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 $\boldsymbol{\Gamma}^\intercal\boldsymbol{\Gamma}$.

Now we observe what happens if we first take the PCA of $\mathbf{X}_{N \times d'}$ and $\mathbf{Y}_{N \times d}$ and then repeat the same exercise. In this case we will take we are looking for the CCA of transformed variables 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 $\mathbf{Y}$. Then, by following the resoning above, the CCA will be obtained from

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

Note that the columns of $\tilde{\mathbf{X}}$ and $\tilde{\mathbf{Y}}$ are orthonormal. If now one considers the transformation introduced and rewrites the above then

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

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

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

If $d = d'$, then

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

and

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

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 $\mathbf{\Lambda}_2^{-1}(\mathbf{Y}\mathbf{Y}^\intercal)$ and $\mathbf{\Lambda}_1^{-1}(\mathbf{X}\mathbf{X}^\intercal)$ respectively. We consider the case $d = d'$ since in the experiments this will be exactly our case. By looking at the obtained results, the PCA-CCA will provide eigenvalues which are extracted by a rescaled matrix, still carrying and considering linear relationship. In practice, this is not enough since the data sets selected contained high non-stationary and non-linear contents, and, therefore, the kPCA-CCA described next is required.

Now, if one consider the non-linear case, one will have

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

$$\begin{aligned}
\underbrace{\mathbf{A}_2}_{N\times p} &= \underbrace{\mathbf{K}_2}_{N\times N}\underbrace{\mathbf{W}_2}_{N\times p} \quad \text{for } \mathbf{W}_2 \text{ kPCs s.t.}\\
&\mathbf{W}_2^\intercal\mathbf{W}_2 = \mathbb{I}_p\\
&\mathbf{W}_2^\intercal\mathbf{K}_2 = \mathbf{\Lambda}_2\mathbf{W}_2^\intercal\\
&\tilde{\mathbf{A}}_{2,i}, \tilde{\mathbf{A}}_{2,j} \text{ are independent}\\
&\mathbf{K}_2 = \mathbf{\Psi}\mathbf{\Psi}^\intercal
\end{aligned}$$

where $\mathbf{\Phi}$ and $\mathbf{\Psi}$ represent the non-linear maps applied to $\mathbf{X}_{N\times d'}$ 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 eigendecomposition. Hence, this time, the CCA will be obtained from

$$\underbrace{\hat{\mathbf{\Gamma}}\hat{\mathbf{\Gamma}}^\intercal}_{p'\times p'} = (\mathbf{A}_1^\intercal\mathbf{A}_1)^{-1}\mathbf{A}_1^\intercal\mathbf{A}_2(\mathbf{A}_2^\intercal\mathbf{A}_2)^{-1}\mathbf{A}_2^\intercal\mathbf{A}_1$$
$$\underbrace{\hat{\mathbf{\Gamma}}^\intercal\hat{\mathbf{\Gamma}}}_{p\times p} = (\mathbf{A}_2^\intercal\mathbf{A}_2)^{-1}\mathbf{A}_2^\intercal\mathbf{A}_1(\mathbf{A}_1^\intercal\mathbf{A}_1)^{-1}\mathbf{A}_1^\intercal\mathbf{A_2}$$

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

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

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

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

If $p = p'$, then

$$\begin{aligned}\underbrace{\hat{\mathbf{\Gamma}}\hat{\mathbf{\Gamma}}^\intercal}_{p'\times p'} &= \mathbf{\Lambda}_1^{-1}\mathbf{\Lambda}_2^{-1}\mathbf{\Lambda}_1(\mathbf{K}_2\mathbf{K}_2^\intercal)\\ &= \mathbf{\Lambda}_2^{-1}(\mathbf{K}_2\mathbf{K}_2^\intercal)\\ \underbrace{\hat{\mathbf{\Gamma}}^\intercal\hat{\mathbf{\Gamma}}}_{p\times p} &= \mathbf{\Lambda}_2^{-1}\mathbf{\Lambda}_1^{-1}\mathbf{\Lambda}_2(\mathbf{K}_1\mathbf{K}_1^\intercal)\\ &= \mathbf{\Lambda}_1^{-1}(\mathbf{K}_1\mathbf{K}_1^\intercal)\end{aligned}$$

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 $\mathbf{\Lambda}_2^{-1}(\mathbf{K}_2\mathbf{K}_2^\intercal)$ and $\mathbf{\Lambda}_1^{-1}(\mathbf{K}_1\mathbf{K}_1^\intercal)$ respectively. This time, it is possible to observe how the employment of the kernel before applying the CCA will be now highly relevant. Indeed, the CCA will extract eigenvalues of a rescaled matrix given by the product of the gram matrix. Hence, the non-linearity and non-stationarity will be faced bu the use of the kernel.

4 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[4].

The construction of these data sets was executed as follows:

1. **Pollution Data:** Sourced from the U.S. Environmental Protection Agency website (`https://www.epa.gov/`), this data set provides a comprehensive view of environmental pollutants in multiple parameters.
2. **Climate Data:** We utilised 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`). This data set offers a daily overview of global climatic conditions.
3. **Green Bonds Data:** This data set was collected through the Bloomberg Terminal, providing comprehensive information on municipal green bonds issued within the state of California in the United States.

Each dataset is submitted to ad hoc data cleaning and feature extraction processes which will produce the desired features necessary for our index construction, i.e. the kPCs. We claim to construct a unique index capturing the contained information of the dataset to aggregate variables in a non-linear fashion. The method to promote such a construction is the KPCA presented in section 2.2.2, providing the set of KPCs that account for non-stationarity and non-linearity of the underlying datasets. In parallel, we will compare these to their linear correspondent representations, i.e. the PCs and show the relevance of the kPCs obtained.

[4]The result of this extensive operation was the creation of three accessible data sets, available at `https://github.com/mcampi111`, which are poised to be invaluable assets for future research endeavours.

Some preliminary steps for the pollution and climate datasets should be considered since they were applied in both cases. Indeed, these datasets require variable collections through specific monitors/stations, which are placed in the counties according to different criteria such as distance from towns/cities, altitude, distance from the sea, etc. Moreover, there might be several constraints in placing a climate or pollution station linked to political, economic, and geographical factors. Understanding such criteria is beyond the main scope of this paper but the reader must be aware of these and can find further insights within the referenced websites below introduced (with respect to each dataset). Figure 25 shows the stations available for the climate and pollution datasets in all US states and California. Note that there is a great deal more in California compared to the rest of the US States, which further incentivises our research in these counties. Two important points should be considered in this regard for our study. The first involves the adopted monitor selection procedure. We claim that major impact of green bonds should be observed in zones with a 50 km radius around cities of at least 250,000 inhabitants. The second aspect is how to relate indices that capture financial information and indices related to pollution or climate. To better understand such a point, consider the financial dataset. The green bonds issued will be categorised according to the geographical area according to the county of issuance. Therefore, there will be no reference in terms of specific cities, but rather to counties. This information drives the study, since we will construct indices at a county level to observe the impact of the green bonds within such areas by considering monitors/stations of pollution and climate variables within a radius of 50 km of major cities. In this way, the most relevant impact can be identified. The CCA description section will highlight this second point in more detail.

Table 2 presented below shows the cities with a population greater than 250,000 inhabitants used to select the climate and pollution monitors. Note that the population number, latitude, and longitude are also presented.

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

Table 2: Table providing the major cities in the US state of California. Particularly, these have a population which is bigger than 250,000 inhabitants. Such cities have been employed for the selection of pollution stations, i.e. the ones falling within a 50 km radius.

The rest of the section is organised as follows: the three data sets are carefully presented. Data pre-processing and extracting the required features for the indices are shown for each. The information carried by the extracted features is described to understand their captured information. Afterwards, the results of the hyperparameter learning procedure for the selection of optimal kPCs are discussed. Once these elements are presented, the following step is to review the results of the linear solution provided by the PCs, which provide a set of linear indices over data sets carrying non-linear and non-stationary features. Afterwards, the results of the kPCs, along with their interpretation for non-linear indices construction, are described. The last part of the section is dedicated to the CCA results over the kPCs.

4.1 The Pollution Data Set

The data set is the pollution data set and is denoted as $\mathbf{X}^1_{N_1 \times D_1}$. As introduced above, D_1 represents the index for observed signals corresponding to carbon dioxide (Co2), nitrous oxide (No2), air quality (AQI), and particular matter 2.5 (PM2.5). Furthermore, the index $N_1 = T_1 \times S$, where T_1 corresponds to the number of timestamps collected for every s location and S the total number of locations, counties in California. The United States Environmental

Protection Agency or the EPA website at `https://www.epa.gov/` provides pollutant stations that collect such variables. Ten years of daily data (2010-2020) have been downloaded by `https://aqs.epa.gov/aqsweb/airdata/download_files.html#Meta`. The polluted stations considered are selected within a maximum distance radius of 50 km from the main cities in Table 2. To further describe the constructed dataset, we first introduce some more notations to proceed.

Denote a pollution variable as $v^m_{i,t,s}$ for $i = 1, \dots, 4$ the index related to the variables of interest, $s = 1, \dots, S$ the index for the county, and $m = 1, \dots, M_{i,s}$ the pollution station where $v^m_{i,t,s}$ has been observed; note that S is the maximum number of counties. Define $t^m_{i,s} = 1, \dots, T^m_{i,s}$ the index for the number of days the i-th variable has been observed in the s-th county and for the m-th station within the radius specified above for that county. Such an index $T^m_{i,s}$ is required since changing for i, s, m, i.e., the observed variables might have different timestamps according to station and county. For example, Co2 in the county of Los Angeles could be observed for 365 days (considering only one year within the 10-year range taken into account and below specified) at station m_1 but only 250 days at station m_2. Alternatively, Co2 might be observed 365 days for a specific year in Los Angeles in a given station but only 250 days in another given station in Santa Clara, etc. Therefore, missing data affects the data in several dimensions. As a result, in the first instance, we extract (for the given variables) all the available stations from the EPA website within a time span of 2010-2020. Subsequently, according to the data preprocessing and wrangling procedure described below, each $T^m_{i,s}$ will be T_1, collected for every location s. The data cleaning and pre-processing procedures are described in the following subsection.

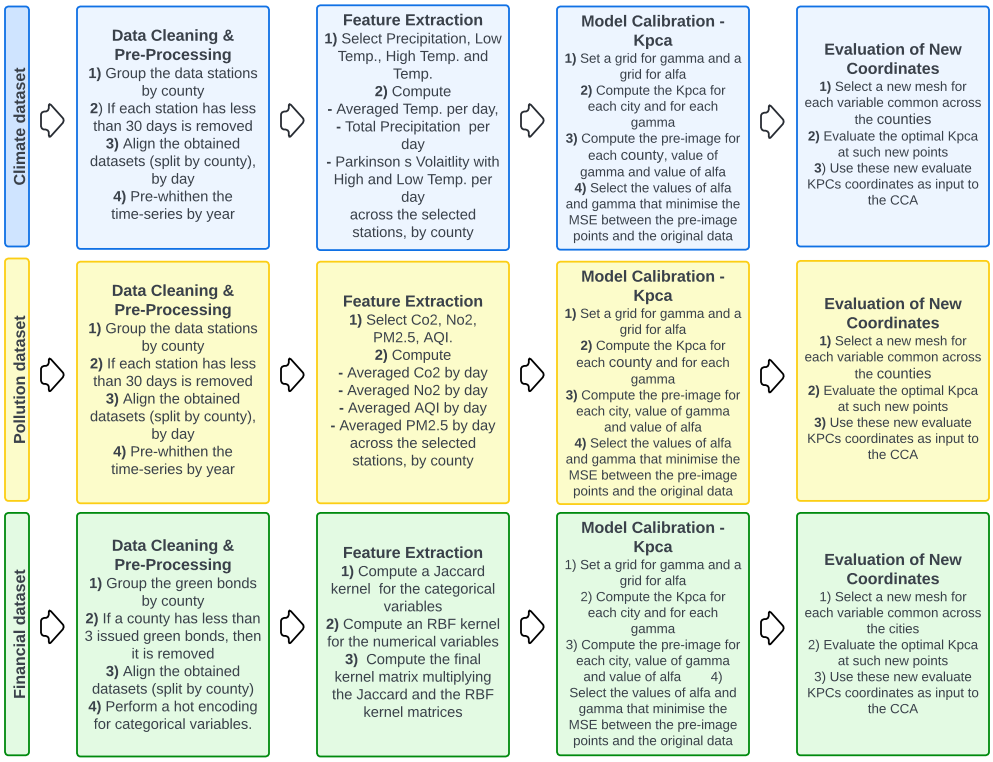


Figure 3: Figure presenting the steps performed on the different datasets.

4.1.1 Data Cleaning and Pre-processing

At this stage, we have four-time series, that is, Co2, No2, AQI, and PM2.5, collected from stations in different counties in California within a radius of 50 km from major cities, that is, cities in California with a population larger than 250,000 inhabitants. The time span for the collection corresponds to 10 years, i.e. 2010-2020. Figure 4 shows four panels where the selected cities are shown in purple and are shown in Table 2. In green, there are the pollution monitors chosen for the study. 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).

Once the stations were selected, we observed how many were within the counties of interest. Figure 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. Observing how the majority fall within Los Angeles, San Diego, and Alameda is possible for all variables. Note that only a subpart from all these counties will be used for the final analysis. Further explanation is provided in the following sections.

Consider now one of the four variables $v^m_{i,t,s}$ and select all observed time series, one for every monitor $m = 1, \ldots, M$ and note that the maximum number of monitors differs between counties; hence we have $M_1, M_2, \ldots, M_S$, as shown in Figure 6. The initial step is to remove the number of stations (across the selected) with timestamps that are less than 30 days. Note that, for every county and every variable, the number of retained stations with less than 1 year timestamps, i.e. less than 365 days (out of 10 year range considered 2010-2020), was never more than 2 stations. Furthermore, in the cases where only one station is retained, then no missingness was present in the considered 10 years time-span. The following step corresponds to, for each county s and for each variable i, aligning the observed values of the stations and computing the mean value per day. This corresponds to $v_{i,t,s} = \frac{1}{M_{i,s}} \sum_{m=1}^{M_{i,s}} v^m_{i,t,s}$. If no timestamp is present for a given station, the mean is taken across the available observations for that given day, ignoring the missing data points. In such a way, we effectively handle missingness and the average per day reflect the data availability accordingly. By considering such a mean value, we define a spatial mean per day directly characterising the variable of interest across the time span of 10 years. Given that there are different number of days per station across both variables and counties, there will be missing days across the averaged $v_{i,t,s}$ (i.e. $T_{i,s}$ differs for both i and s). Therefore, we fit a spline through the missing data. At this point, the same number of timestamps is available across every county spatial average; therefore, there are T_1 final timestamps available. The final step corresponds to re-whitening the obtained time series $v_{i,t,s}$. Note that the diagram given in Figure 5 summarises it. As result, we have four-time series for every county, i.e. averaged Co2, averaged No2, averaged PM 2.5 and averaged AQI derived for each county as a measure synthesising all the stations present in that specific county. The following subsection will describe the interpretation behind such ad hoc constructed features and their utilisation.

4.1.2 Pollution Extracted Features

These four engineered features will capture the variability of AQI, CO2, NO2 and PM2.5 through a spatial average of the considered pollution monitors over a time span of 10 years. Hence, these are expected to be highly non-stationary, possibly carrying a seasonal component depending on the California County considered. Fig. 7 presents four panels, each showing a heatmap of California. The top panels show heatmaps of No2 (left) and Co2 (right), while the bottom panels show PM2.5 (left) and AQI (right), respectively. Note that each county of each plot is represented by a unique colour, i.e. a unique number, for the variable of interest. This results from an average of that variable across the ten-year time series. It is possible to observe how the averaged No2, Co2 and AQI significantly vary across the different counties. In the case of PM2.5, a less substantial variability characterises California counties.

Appendix C presents Fig. 26 in which we provide a set of plots for every variable, but this time, we have one plot per quarter per variable, thus four yearly quarters for No2, Co2, AQI and PM2.5. These figures compute the average across the time of every quarter. In such a way, more variability can be observed across the counties and the quarters. Fig. 8 provides boxplots of the averaged variables by quarters to further understand the high variations carried in the engineered features. Thus, each boxplot is constructed with 10 points of that quarter and that county. There are sixteen panels, one for every engineered feature and every yearly quarter, i.e. Q1, Q2, Q3 and Q4. the top rows represent boxplots of AQI, and the second row shows panels for Co2, followed by No2 and PM2.5, respectively. Observing how great variability comes from the first three mentioned variables across all quarters is possible. PM2.5, instead, appears to change way less within these counties, suggesting that these data are driven by variability within AQI, Co2 and No2.

4.2 The Climate Data Set

The second dataset is the climate dataset and is denoted as $\mathbf{X}^2_{N_2 \times D_2}$. For the pollution data set, D_2 represents the index for the observed signals, corresponding to the mean temperature (.1 Fahrenheit), the maximum and minimum temperatures (.1 Fahrenheit), and the total amount of precipitation (.01 inches). For simplicity, we denote these variables as MT (mean temperature), Ht (maximum temperature), Lt (minimum temperature), and PRC (total amount of precipitation), respectively. These signals are observed and collected daily, within a time range of 10 years, 2010-2020. The $N_2 = T_2 \times S$ corresponds to the number of timestamps collected for every s location and the S total number of California counties. The Global Surface Summary of the Day (GSOD) dataset provides data for the climate stations that contain the variables of interest. This 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. Regarding the pollution data set, the climate stations are selected within a radius of 50 km maximum distance from the major cities in Table 2.

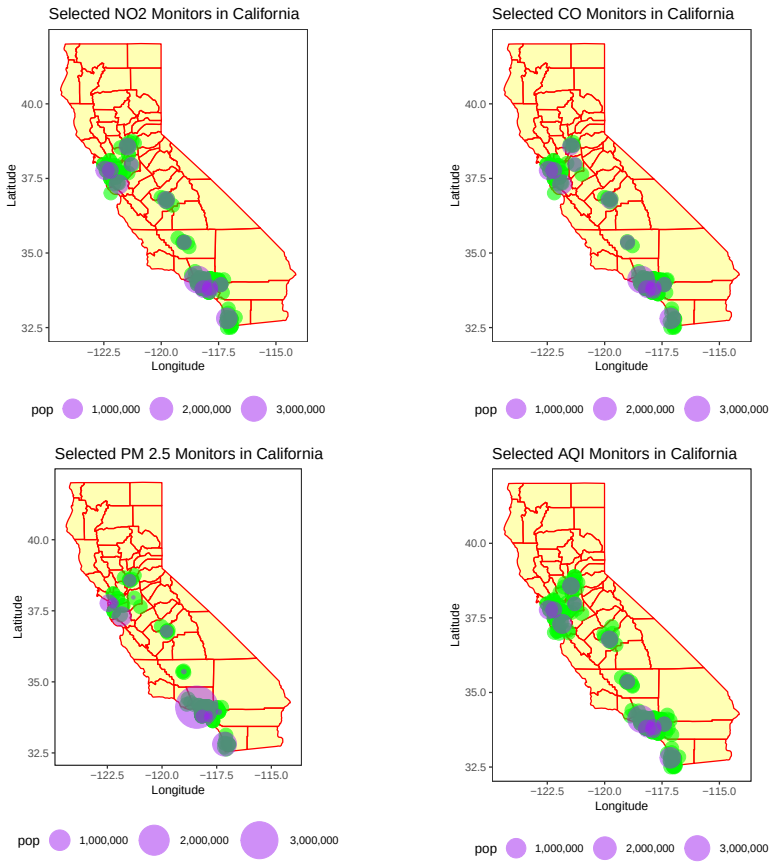


Figure 4: Selected pollution monitors for collecting No2, Co2, PM2.5 and AQI. In purple major cities of California given in Table 2. In green the selected monitors which fall within a radius of 50 km around such cities.

The next part proceeds as for the pollution dataset: first, we consider a climate variable as $u^l_{j,t,s}$ for $j = 1, \dots, 4$ the index related to the variables of interest, $s = 1, \dots, S$ the index for the county, and $l = 1, \dots, L_{j,s}$ the climate station where $u^l_{j,t,s}$ has been observed. Note that S is the maximum number of counties. Define $t^m_{j,s} = 1, \dots, T^l_{j,s}$ the index for the number of days that the variable j-th variable has been observed in the s-th county and for the l-th station within the radius specified above for that county. Again, $T^l_{j,s}$ is required since changing for j, s, l, i.e., the observed variables might have different timestamps according to station and county. Hence, data are missing across both the temporal and the spatial dimensions. First, we extract (for the given variables) all the available stations from the GSOD website in a time span of 2010-2020. Subsequently, according to the data pre-processing and wrangling procedure described below, each $T^l_{i,s}$ will be T_2, collected for each location s. The data cleaning and pre-processing procedures are described in the following Subsection.

4.2.1 Data Cleaning and Pre-processing

We end up with four-time series, i.e. MT, Ht, Lt and PRC, denoted $u^l_{1,t,s}$, $u^l_{2,t,s}$, $u^l_{3,t,s}$ and $u^l_{4,t,s}$ respectively. These are collected from stations in different counties in California within a radius of 50 km from cities with a population bigger than 250,000 inhabitants for a time span ranging between 2010-2020. Figure 9 shows two different panels. The left plot represents the selected cities in purple (as given in table 2) and in green the selected climate monitors used for the study. The right panel shows a barplot with the number of stations within each county. As for the pollution variables, a great deal of monitors fall into the counties of Los Angeles and San Diego.

Consider one of the four climate variables $u^l_{j,t,s}$ and select all observed time series, one for every monitor $l = 1, \dots, L_s$ and note that the maximum number of monitors differs between counties; hence we have $L_1, L_2, \dots, L_S$, as shown

in Figure 9. The procedure of data cleaning goes as for the pollution data set. Therefore, at first, we remove the ones with less than 30 days in total across the pre-selected stations. In this regard, for every county and every variable, the number of retained stations with less than one-year timestamps, i.e. less than 365 days (out of the ten years range considered), was never more than three stations. In the cases where only one station is retained at the final, no missingness was present in the ten years. After this stage, we apply different data-wrangling procedures depending on each variable and define various climate features described in the following paragraph.

4.2.2 Climate Feature Engineering

Different features are now extracted and engineered by working with the four climate time series taken into account. Consider first Mt or $u^l_{2,t,s}$ and mt $u^l_{3,t,s}$ temperatures. For each county s and every station l, we fit a spline were missing data are present generating $\hat{u}^l_{2,t,s}$ and $\hat{u}^l_{3,t,s}$. Afterwards, the following volatility estimator is considered

$$\sigma^l_{t,s} = \sqrt{\frac{1}{4n \times \log 2} \sum_{t=1}^{n} \left(\log \frac{\hat{u}^l_{2,t,s}}{\hat{u}^l_{3,t,s}} \right)^2} \tag{43}$$

Such volatility estimator is the Parkinson volatility [47] and captures information on high and low temperatures in each station per location. Once we obtain the Parkinson volatility per every location s and stations L, we compute an average of such a quantity per day. This is done within each location, i.e. each county. We will therefore end up with one vector per location summarising this volatility per day. We compute the average across the stations per location per day for the mean temperature. In contrast, we calculate the total precipitation per day (hence, we sum up the rainfall recorded within each station of every location by day). Figure 10 summarises these feature engineering methods.

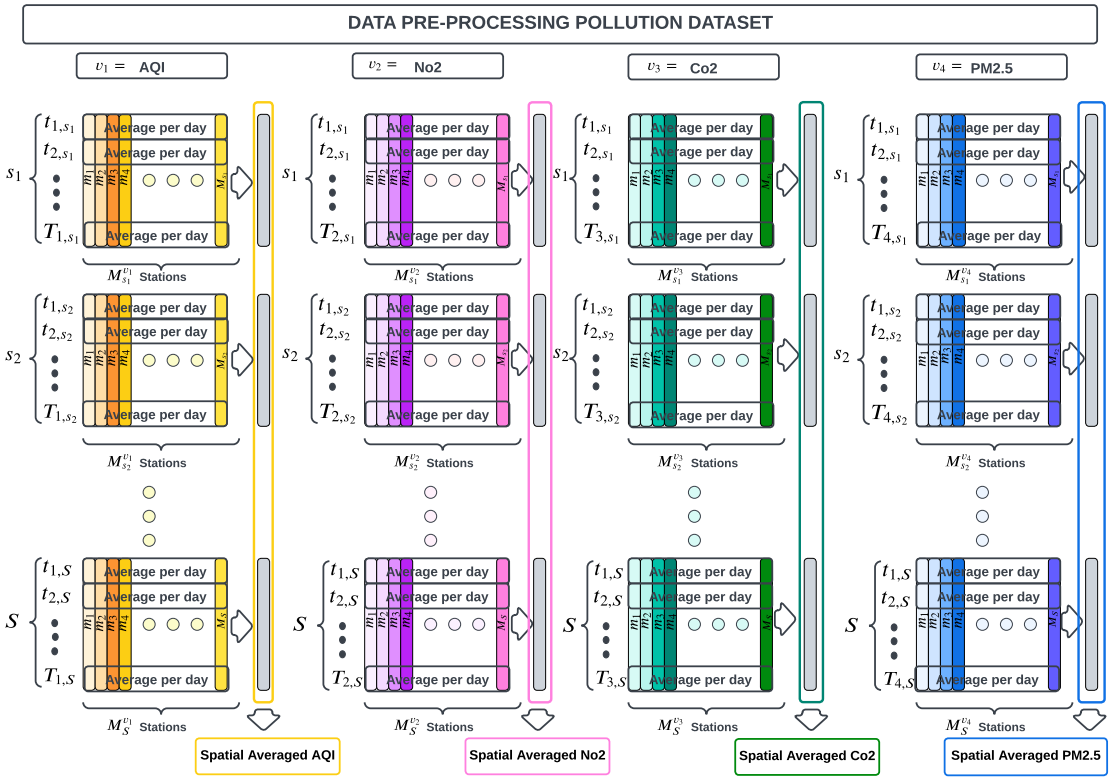


Figure 5: Figure presenting the diagram of the data cleaning and pre-processing of the pollution dataset. The time series of the stations of every location and every variable are the ones which recorded at least 30 days of the variable of interest.

As for the pollution data set, we present two figures. Fig. 11 show heatmaps of the counties of California, one for every feature, i.e. averaged mean temperature (left), averaged total precipitation (middle) and averaged volatility

(right), representative of maximum and minimum temperatures. These features are averaged in these heatmaps because each county is coloured according to a unique colour corresponding to an average of that feature across ten years of time series. Significant variability can be identified within each of the panels across the counties.

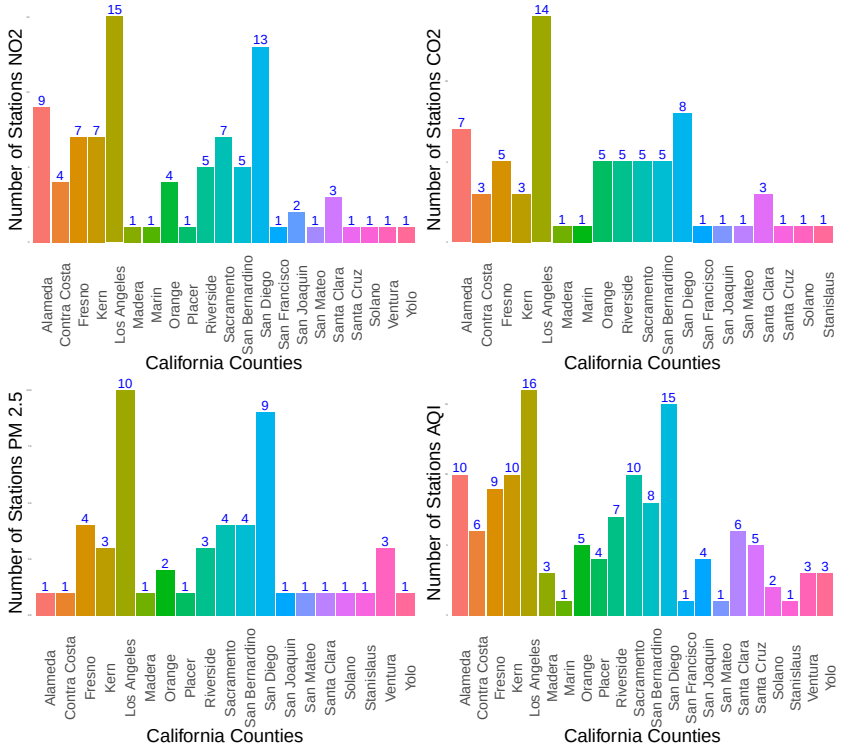


Figure 6: Barplots showing the number of selected stations within each county of interest. Note that the top panels refer to No2 and Co2 (left and right, respectively), while the bottom panels refer to PM2.5 and AQI (left and right, respectively). The x-axis shows the different California Counties (ordered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

Appendix C provides Fig. 27, which shows equivalent heatmaps with averaged features by quarters rather than over the entire ten years. Therefore, there will be four panels for every considered feature. Such a figure shows how significant the variability of these quantities is and how it changes across time and county. The second figure corresponds to Fig. 12 in which boxplots of the averaged features across the quarters are provided. Hence, before discussing the obtained variability, note that each boxplot is representative of ten points, which are the average of that quantity within each of the quarters of the ten considered years. The top rows show averaged mean temperature across the quarters for every county. It is possible to observe how high variation is present during every quarter. A different situation is instead found for total precipitation and average volatility. Indeed, these two features appear to change much more during the first and last quarters while showing lower variations within the year's second and third quarters. Such an effect is strictly related to the seasonal component of such time series.

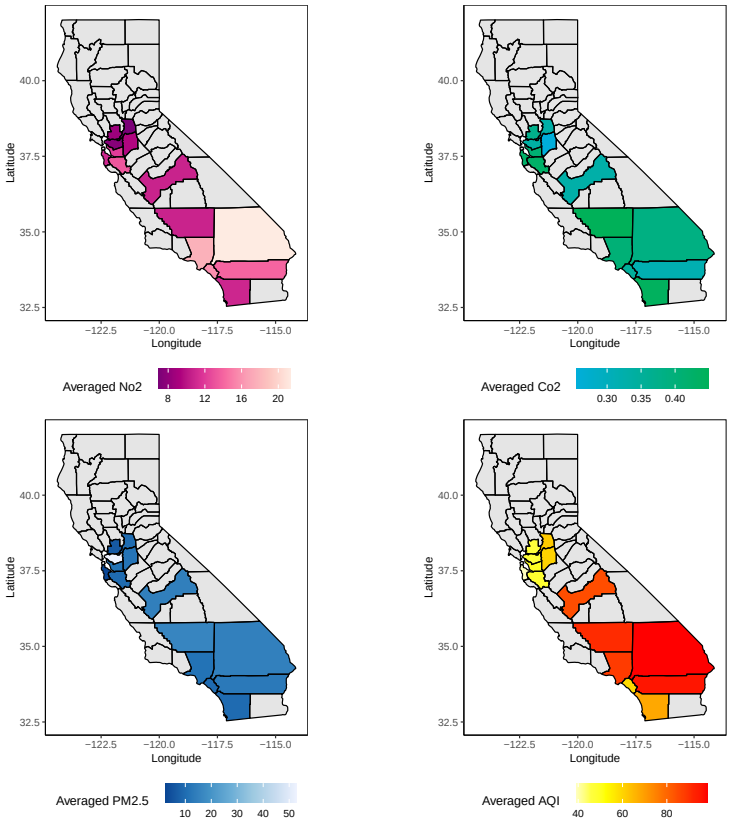


Figure 7: Heatmaps of the engineered pollution features averaged across 10 years by county. In the top panels, there are averaged No2 and Co2 (left and right respectively), while the bottom panels show averaged PM2.5 and AQI (left and right respectively).

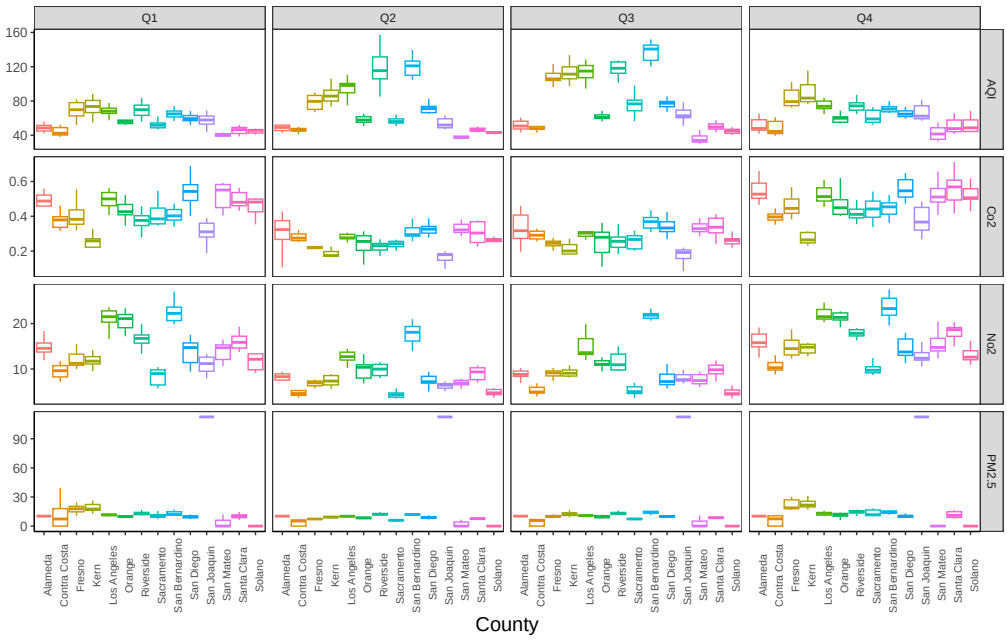


Figure 8: Boxplots of the pollution-engineered features, further averaged by yearly quarters and county. Hence, each boxplot is representative of 10 points, where that variable has been averaged across the quarter and the county.

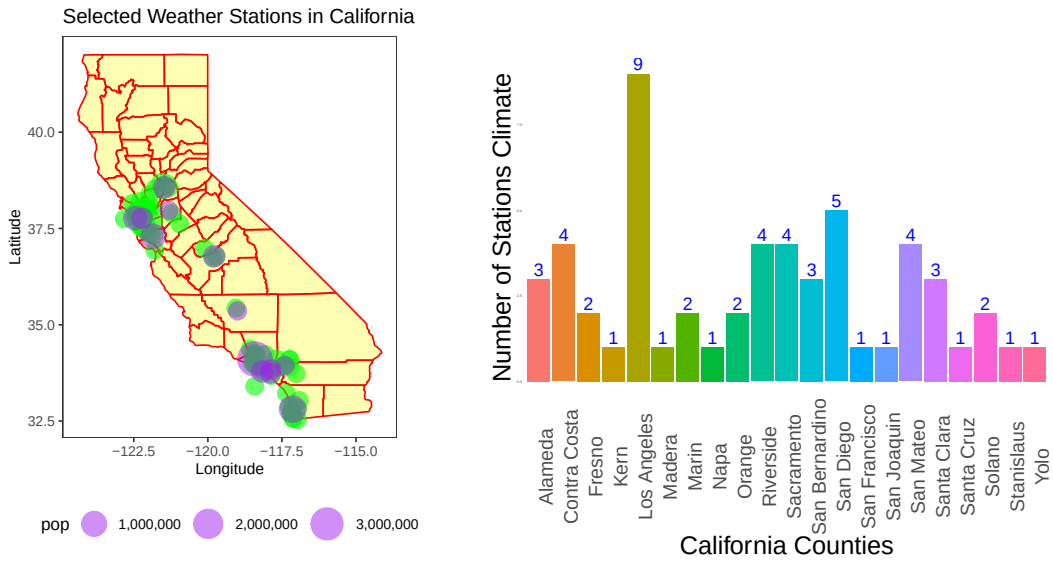


Figure 9: Left panel: map of the selected weather stations in the State of California. In purple are the given counties, while in green are the weather stations around the counties. Right panel: number of stations per counties.

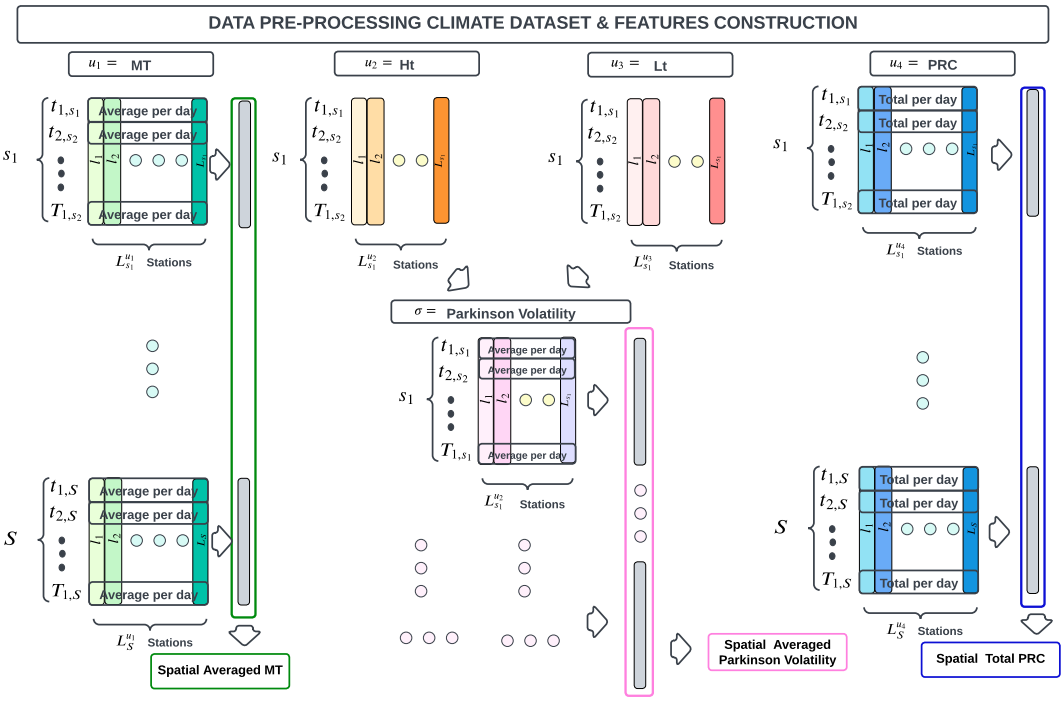


Figure 10: Figure presenting the data cleaning and pre-processing diagram of the climate dataset. The time series of the stations of every location and every variable are the ones which recorded at least 30 days of the variable of interest.

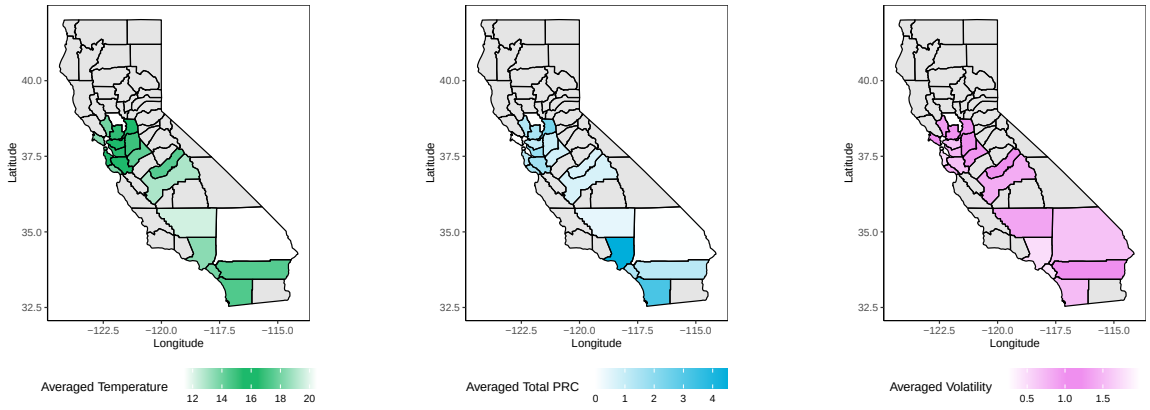


Figure 11: Heatmaps of the engineered pollution features averaged across ten years by county. From left to right, the panels show averaged temperature, averaged total precipitation and averaged volatility. We use the term averaged for each feature, meaning that the plotted value for every county is an average across the ten years.

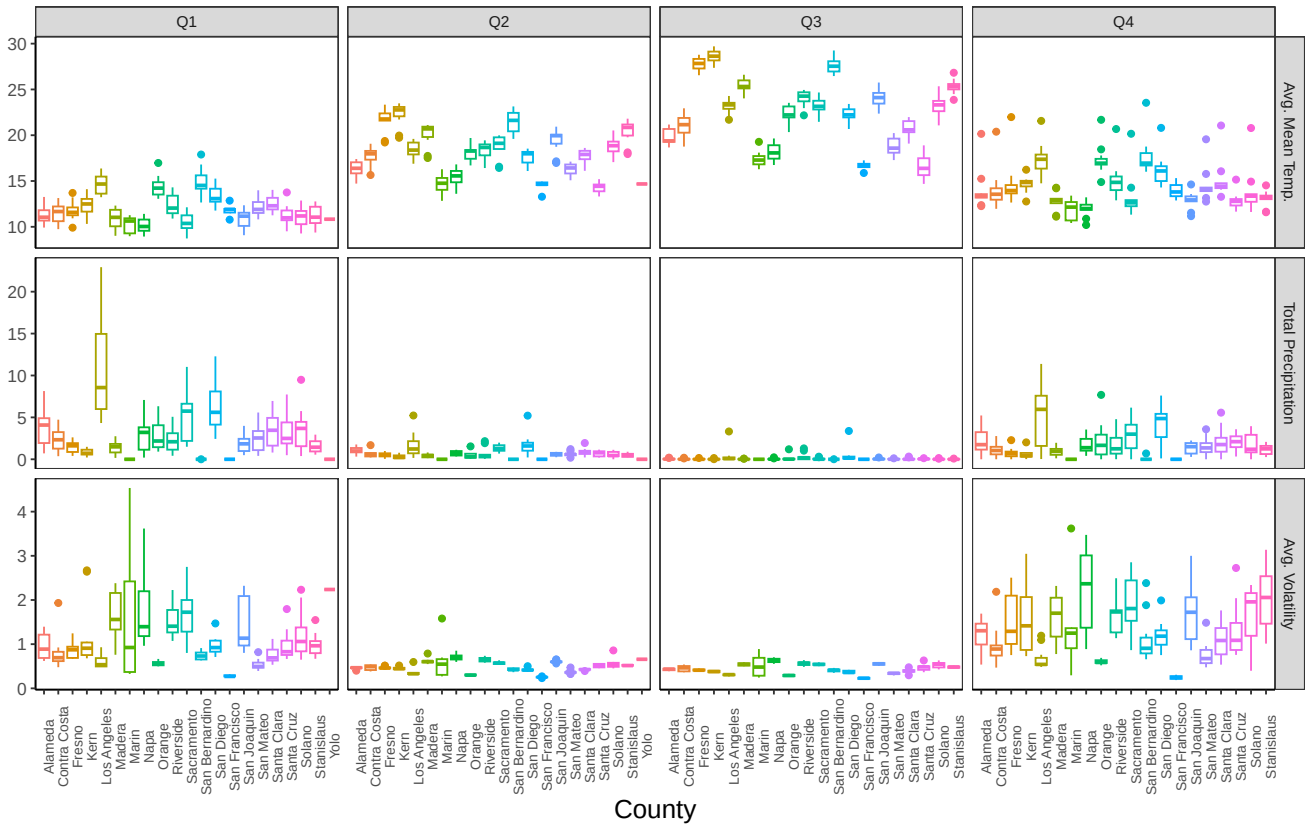


Figure 12: Boxplots of the climate-engineered features, further averaged by yearly quarters and county. Hence, each boxplot is representative of 10 points, where that variable has been averaged across the quarter and the county.

4.3 The Financial Data Set

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. Several characteristics of green bonds may enhance environmental outcomes, particularly in the context of municipal bonds.

These characteristics of municipal green bonds - their issuance size, maturity, the amount issued, yield at issue, coupon, spread, credit risk, and others - are not only financial attributes. They also embody crucial dynamics that shape the market and its potential for environmental impact. Their importance lies in their ability to enhance the market's appeal to diverse investors, thereby catalysing growth and ensuring the market's liquidity. Liquidity in the green bond market allows for more active trading and facilitates price discovery, further increasing market attractiveness. A robust, diverse green bond market not only spreads risk but also encourages innovation, paving the way for improved environmental outcomes. Essentially, the richness and breadth of these bond attributes lay the foundation for a mature, liquid, and transparent green bond market. As this 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 blend of these attributes plays a crucial role in amplifying the environmental impact of green bonds, particularly in the municipal context.

We, therefore, aim to identify the variation within such data to identify the impact of these instruments and quantify their carried variations within the financial market. More formally, we define the third data set considered as the financial data set and denoted as $\mathbf{X}^3_{N_3 \times D_3}$. As above, D_3 represents the index for the observed signals or variables. These will be introduced in the below subsections. In the formerly described data set, the index N referred to the total number of days observed, i.e., ten years of data for the climate and pollution cases (from 2010 to 2020). The financial data corresponds to green bonds issued by munies in the US State of California. Therefore, N_3 corresponds to the number of green bonds issued in this county. The considered time span for issuing is five years, from 2015 to 2020. Figure 13 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. Note that, as for the other data sets, Appendix D presents Figure 28, which provides the number of issuers of green bonds regarding munies all over the US. It is possible to observe how California represents the State where most issuers are present and, therefore, the one studied in this work. The extraction of green bond data for munies has been conducted through the Bloomberg Portal. This exercise requires an advanced screening for the selection of the desired data and avoids the use of fields which are not redundant in the analysis. The screening procedure is presented in the section below so that reproducible results can be achieved. Note that we conducted the screening procedure at a US level and afterwards subselected the one for the state of California. Another critical point is that this data set required ad hoc constructed solutions for the kernel PCA extraction. This is because the collected variables describing the green bonds were numerical, categorical or dates. Cleaning procedures and hot encoding were employed to treat such differences. Such processes are presented in the sections below. Hence, this subsection firstly reviews the financial data extraction from Bloomberg, explaining the criteria step by step. Afterwards, the collected financial variables describing the green bonds are presented. Finally, the data cleaning procedure and the kernel construction for this mixed-type data are discussed.

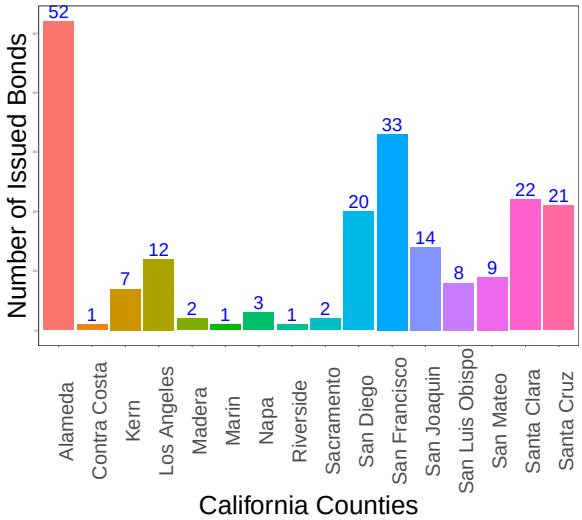


Figure 13: Number of green bonds issued in California. Note that the total corresponds to 208. However, in the analysis, due to some data cleaning criteria and kernel construction, the final number will be 167. Details are given in the subsections below.

4.3.1 Financial data extraction from Bloomberg

To identify eligible municipal green bonds and conventional bonds in the primary market, we used Bloomberg's search function 'SRCH' to screen both types of bonds. This market screening function allows users to create customized 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, 144A, and STRIPs), including non-Bloomberg-verified bonds, and including strips (loans). None of these options were selected, and when tested, they did not alter the final count of assets.

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 toward 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.

Tables 3, 4, and 5 summarize the filter criteria and the resulting number of bonds that fit the criteria within the previously filtered results. The "Matches" column displays the number of bonds that fit the criteria within the previously filtered results. 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, outstanding amount, and more. Table 5 displays the selection criteria for Municipal green bonds (as of October 19, 2020) without optionality, resulting in 3,436 matches. The selection process took several steps and involved refining the dataset to ensure the most relevant and accurate bonds were included for analysis. This comprehensive approach allowed us to gather valuable insights into the green bond market and examine the specific attributes of the selected bonds that could influence their market value and impact.

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 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
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[5]	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
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

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

4.3.2 Financial Variables

Once the screening procedures were complete, we obtained 208 green bonds issued in the US State of California, for which a set of variables have been collected. In practice, we will consider only a sub-selection of the counties (9 in total below specified) and due to data cleaning and kernel construction procedure, this will result in 167 green bonds in total. Table 6 shows information about the collected variables, i.e. the name, a brief description and the data type. From there, we will derive the kPCs describing such a dataset's nonlinear and non-stationary variability. This is a daily challenge for practitioners, and therefore, the first aim of this work is to describe the variation through an indicator dealing with different data structures and high rates of time or spatial variations. To better understand the information captured through the constructed kPC index, first, we review what these variables represent in the green bond context to analyse where the variation might come from.

Each green bond will be identified by a unique code called "CUSIP", allowing the extraction of all collected variables. We identified information on the municipal bond's maturity size, issuance date, and maturity date. Green bonds are instruments specifically issued to finance environmentally friendly or sustainable projects, whose structure can vary and can be issued in different forms, including conventional and zero coupon bonds. Most green bonds are structured as coupon bonds, which means they pay periodic interest payments (coupons) to bondholders over the bond's term.

Coupon payments may be fixed or floating, depending on the specific terms of the bond. At maturity, the bondholder receives the final principal payment. However, some green bonds may also be structured as zero-coupon bonds. Zero coupon bonds do not pay periodic interest payments like coupon bonds. Instead, they are issued at a discount to their face value and mature at their full face value, with the investor earning the difference between the purchase price and the face value as the return. Both coupon bonds and zero-coupon bonds can be used as vehicles for green bond issuance, depending on the preferences of the issuer and the market conditions. The choice of bond structure, including whether it is a zero-coupon bond, will depend on factors such as the issuer's funding needs, market demand, and investor preferences. The coupon, paired with the spread over risk-free rates, often influences investors' attractiveness of green bonds. A higher coupon or spread can entice green bonds to investors seeking higher returns, potentially leading to more capital flowing into green projects.

The maturity size of a green bond refers to the time until the bond reaches its maturity date. It represents the period during which the bondholder will hold the bond before receiving the final principal payment. The maturity size of a green bond can vary and is determined by the issuer at the time of issuance. It is typically specified in the bond offering documents, such as the prospectus or official statement, and can range from a few years to several decades. The choice of maturity size for a green bond depends on various factors, including the nature of the financed project, the issuer's funding needs, and market conditions. Green bonds issued to fund shorter-term projects, such as energy efficiency retrofits or waste management initiatives, may have shorter maturity sizes, ranging from 3 to 10 years. On the other hand, bonds financing longer-term projects like renewable energy infrastructure or sustainable transportation projects may have maturity sizes of 10, 20, or even 30 years. Offering a range of maturities can attract a broader array of investors, each with different investment horizons, and create a yield curve, an essential tool for fixed-income investors. A more complete green bond curve would aid price discovery and improve the market's efficiency.

The issue price of a green bond refers to the price at which the bond is initially offered or sold to investors when it is issued by the issuer. It represents the initial cost or value at which the bond is priced and made available for purchase in the primary market. As for other bond markets, this price is determined by several factors, including market conditions, investor demand, the creditworthiness of the issuer, and the specific terms of the bond. The issuer, often in consultation with underwriters and financial advisors, sets the issue price on the basis of these considerations.

In most cases, the issue price of a bond is set at par, which means that the bond is issued at its face value. However, bonds can also be issued at a premium (above par) or at a discount (below par) depending on the prevailing market conditions and the appetite of investors. Once the bond is issued and begins trading on the secondary market, its price may fluctuate according to various factors, such as market interest rate changes, credit rating updates, and supply and demand dynamics. Investors need to understand the issue price as it represents the starting point for evaluating the bond's performance, potential returns, and overall value.

The amount issued refers to the total value or size of the green bonds that are issued by the issuer. It represents the aggregate principal amount of bonds that are sold to investors at the time of issuance. It can vary depending on the specific offering and the funding needs of the issuer. It is typically disclosed in the bond offering documents and represents the total capital raised through the issuance of green bonds. The yield at issue, also known as the Initial Yield or Coupon Yield, refers to the effective interest rate or yield at which the green bonds are initially offered to investors. It represents the return investors will receive in the form of periodic coupon payments relative to the bond's price. The Yield at Issue is typically stated as an annual percentage rate (APR) and is set by the issuer based on market conditions, credit risk, and the specific terms of the bond. It is designed to make the bond competitive with prevailing interest rates in the market. These two quantities are important considerations for investors and provide insight into the size of the green bond offering and the return potential at the time of the bond's issuance. A lower yield at issue could make green bonds more attractive to issuers, potentially increasing the number of green projects funded. These terms and other features, such as maturity, credit rating, and use of proceeds, are typically disclosed in the bond prospectus or official statement, allowing investors to make informed decisions about investing in green bonds.

The BID OAS (Option-Adjusted Spread) represents the number of basis points the spot curve would have to shift for the present value of the cash flows to equal thesecuritys prices. Thus, it provides valuable information that places a green bond in a context concerning its matter on the market. It measures the bond's relative value and is commonly used in bond pricing and valuation. It considers factors such as the bond's coupon rate, maturity, call features, and market conditions to determine its fair value. For green bonds specifically, the BID OAS reflects the spread over the risk-free benchmark rate associated with the green bond's specific credit risk and any embedded options, such as call options or put options, that may be present in the bond's terms.

The Mid Macaulay duration measures the weighted average time until a bond's cash flows are received, considering the present value of those cash flows. It helps to assess the bond's interest rate risk, with a higher duration indicating a greater sensitivity to changes in interest rates.

The spread at issuance to worst describes the spread between a bond's initial yield at the time of issuance and its yield under its worst-case scenario. It measures the additional yield an investor demands for holding the bond, considering the potential credit risk and other factors that could affect the bond's value. The worst-case scenario considers factors such as a credit rating downgrade, default risk, or adverse events that may impact the bond's performance. It is an important consideration for investors as it provides insight into the compensation they receive for taking on potential risks associated with the green bond. A wider spread indicates higher perceived risk, while a narrower spread indicates lower perceived risk.

Figure 14 plots the different numerical variables considered in the kPCs extraction for every county considered. It is possible to observe how each variable varies highly across and within the different counties. Many of the boxplots are widely spread, suggesting a high level of variations. Hence, a cross-correlation analysis dealing with non-stationarity, non-linearity and different structured data, given the presence of categorical data in this data, is strictly required in this work.

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 Issued	Cumulative amount issued from the original security pricing date through to the current date for debt securities	Numerical
Coupon	Current interest rate of the security. For bonds with reset compounding structures, this will return the estimated annualized daily reset compounding structures, this will return the estimated annualized rate for coupon cash flow calculations for the corresponding settlement date	Numerical
Issue Date	Date the security is issued	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

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.

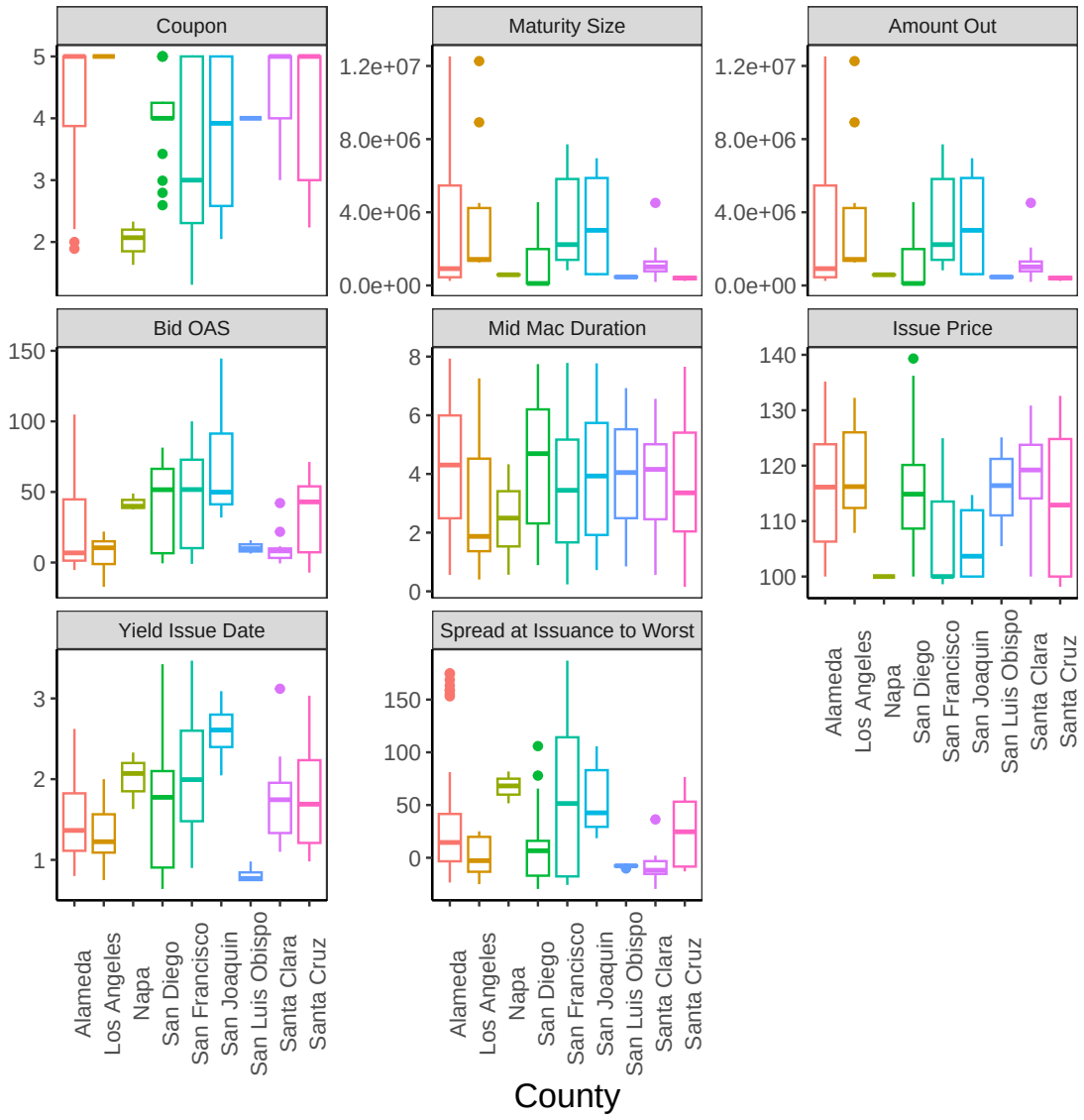


Figure 14: Boxplots of the numerical variables used for the financial data set. The x-axis shows the considered Counties while the y-axis the range of the values for every variable.

Figure 15 shows 4 panels. The left-top panel describes Muni Issue Type and the right-top panel refers to Muni Source. By focusing on the Muni Issue Type, it is possible to observe that 127 bonds are revenue bonds, followed by General Obligation UNLTD. The other Muni Issue Types are Tax Allocation (18 bonds), Special Tax (16 bonds), Certificate Participation (13 bonds), General Obligation LTD (4 bonds) and Special Assessment (1 bond). Regarding the Muni Source, the ones with the majority of issued bonds correspond to Ad Valorem Property Tax (67 bonds), Lease (Abatement) (46 bonds), and Water Revenue (32 bonds). The bottom panels refer to Issuer Industry and the Bloomberg Issuer 5-Year Credit Risk, left and right, respectively. The Issuer Industry mainly has 3 bigger classes being GEN (99 bonds), GOB (49 bonds) and Water (42 bonds). The following with relatively big number of issued bonds are SCD (18 bonds), MEL (16 bonds) and POL (12 bonds). Regarding the Bloomberg Issuer 5-Year Credit Risk, the issued bond is classified into eight different risks. 77 bonds are IG4, 65 are IG3, 30 IG1 and most of the remaining are in the classes of IG2 and IG5. A high variation will be expected Given the different distributions of bonds across these variables. Appendix D presents Fig. 29, which is a more refined version of Fig. 15 but split by county, to observe these attributes distributions within each of the considered location.

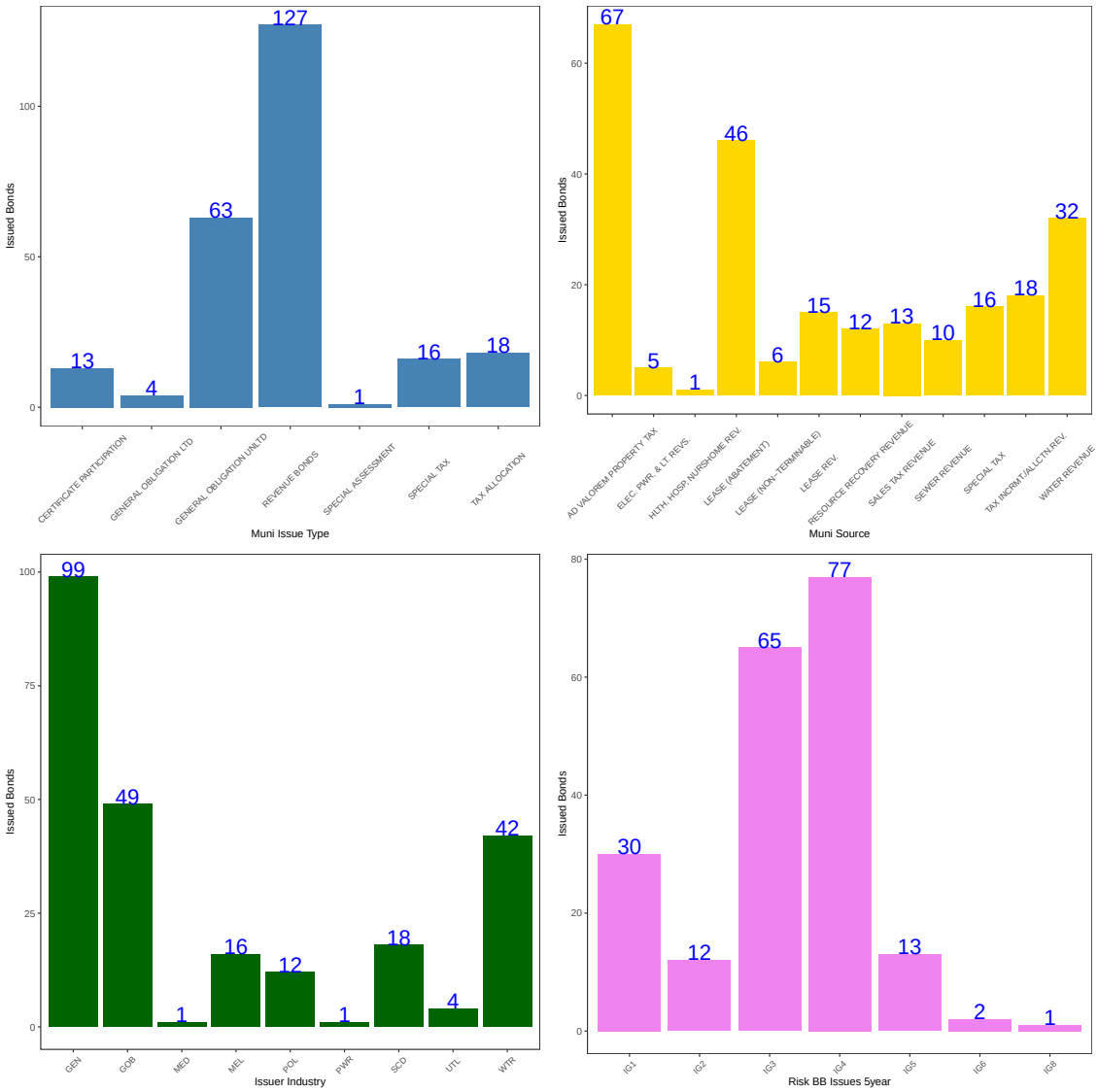


Figure 15: Barplots of some of the used financial variables. The top plots refer to Muni Issue Type (left) and Muni Source (right); while, the bottom panels refer to Issuer Industry (left) and Bloomberg Issuer 5-Year Credit Risk (right).

4.3.3 Data Cleaning, Hot Encoding and Kernel Construction

In this section, we review the data cleaning and hot encoding applied to the financial data set. Afterwards, we present how the kernel matrices between categorical, numerical and date data have been constructed for the kPCA.

First, each California county with less than three green bonds issued will be removed. After that, we separate the categorical and date variables from the numerical ones, as given in Table 6. In this way, the hot encoding can then be performed. Note that we end up with 167 green bonds in total, which will then be split per county location of the issuer, as shown in Fig. 13. 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 [48] 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.

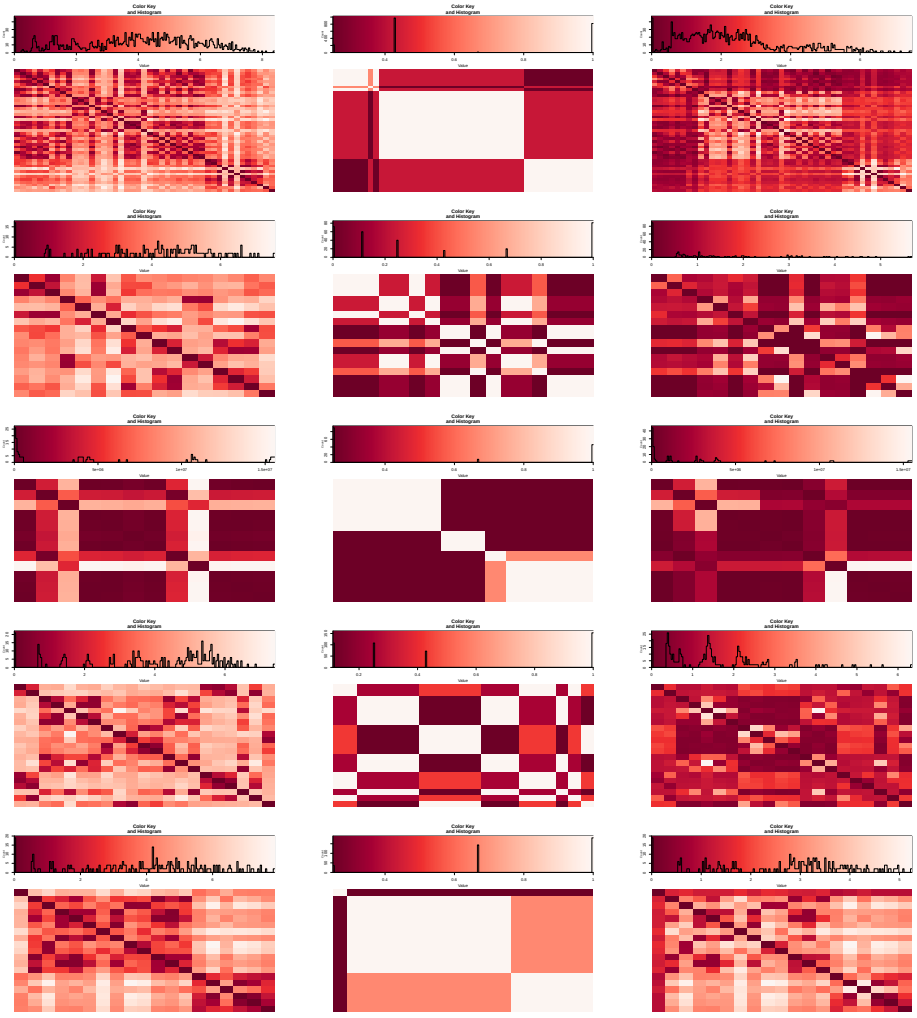


Figure 16: Heatmaps of the Gram matrices computed on the financial variables of the counties Alameda, San Francisco, Los Angeles, Santa Cruz, San Diego. The left panels correspond to the radial basis kernel Gram Matrix computed on the numerical and encoded date financial attributes; the middle panels represent the Jaccard kernel Gram Matrix calculated on the categorical encoded variables. The right panels correspond to the sum of the two matrices that provide the final matrix for all financial variables. The same procedure has been applied to each county.

Once the encoding of all the variables is performed, we can construct a kernel matrix from which we can extract the kPCA. Firstly, we group the variables into separate groups, i.e. numerical and encoded date variables will be in one group, and the encoded categorical variables will be in the second one. Afterwards, we will construct kernel matrices on these two sets of variables by counties. For the first group, we use the radial basis function; for the second group, we will use the Jaccard distance, thus computing the Jaccard kernel matrices. Fig. 16 shows some examples of the matrices computed on the numerical and date attributes (left panels) and the one on the categorical ones (middle panels) 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. The examples for Alameda, San Francisco, Los Angeles, Santa Cruz, San Diego are given in the right panels of Fig. 16. Note that we computed such Gram matrices for every county, and then, once the final matrices are obtained, we compute kPCA on these.

To further observe the different contribution of the categorical variables, Appendix E show alignments of empirical covariance matrices of different combinations of categorical variables with the covariance matrices of the original data, by county. The idea is to observe which categorical variables contribute the most to the variation of the extracted kPCs. More details on the alignments employed are given in Subsection 4.7, since this measure is also used in this work for other purposes.

4.4 Results of Hyperparameter Learning

Once the kPCA is computed on each data set, the last part of the procedure for each considered data set corresponds to identifying the optimal kernel parameters, i.e., the γ parameter of the radial basis kernel function. This is obtained through the pre-image method and the hyperparameter learning algorithm. In practice, we compute the distances between the obtained pre-image for every tested hyperparameter and every data set and then minimise 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.

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

Table 7: Table describing the optimal γ parameters of the RBF kernel as given in Equation (38). The procedure for identifying such final hyperparameters is summarised in algorithm 1. As explained, the Kpca is conducted at a county level, hence the first column present 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 4). The columns represents the three different data sets, i.e. financial data set, pollution data set and climate data set.

4.5 Linear Solution: PCA

As the PCA-CCA is our benchmark model for the implemented kPCA-CCA, this subsection aims to analyse the explanatory power of the PCs. This is achieved by looking at the percentage of captured variance to assess which PCs carry the greatest data variability. The qualitative and quantitative analyses are then conducted, relying on different visual representations of the PCs. Note that we extract the PCs (and the kPCs) for pollution and climate over time, i.e. these are ten-year time series. We constructed a set of plots averaging the PCs by counties and quarters, i.e. Q1, Q2, Q3, and Q4 of the ten years (rather than plotting them on the entire period). The climate and pollution variables are highly seasonal; therefore, a visualisation across the whole time would make patterns over time more difficult to be identified. Since the PCs of the financial data are extracted on attributes that do not carry any time evolution, in this case, we observe the first three PCs by county only. In the qualitative analysis, we used a penalised spline interpolating the averaged PCs across quarters of the pollution and climate data sets and the original PCs characterising the financial data. We plot the standard error (confidence interval) around the fitted spline. In the quantitative analysis, we observe boxplots of these same quantities, i.e. PCs by counties and quarters for pollution and climate data and by county for the PCs of the financial data.

When observing these plots, one should recall that there is a notional difference between process uncertainty (or systematic variance) and statistical uncertainty. The former depends on the inherent process variation due to the non-stationarity of the studied signals. Ideally, if a signal is stationary in all higher-order moments, then the bases chosen to model it (PCs or kPCs) would be obtained exactly. Nevertheless, if the signal is non-stationary, whichever selected model solution (linear or non-linear), would always lead to certain misspecification (due to several model parameters, model selection, etc). This misspecification could affect the results of the spline fitting, leading to higher standard

deviation and larger confidence intervals. On the other hand, statistical uncertainty can be induced if the sample size is not large enough, producing a larger standard deviation of the fitted spline or a wider confidence interval. Accordingly, the qualitative and quantitative analyses must account for these two uncertainties.

Table 8 shows the percentage of variance of the three PCs of each data set, expressed between 0 and 1. We highlighted the percentages that are superior (or equal) to 50%. In the pollution and climate data sets, the first PC captures more than 50% of the underlying variability, with very few exceptions. Moving across the PCs, the second detects around 20% of variation, and the third PC around 10%. In the case of the financial data set, the only county in which PC1 detects 50% of the variability is Napa, which has very few samples, that is, green bonds issued. For the rest of the counties, the first three PCs always detect less than 30% of data variability, suggesting low PC performance. Thus, the PCs of pollution and climate show good explanatory power when the first PC, i.e. a rank-one linear approximation matrix, is considered. These good performances of a linear model indicate that the caught variability is well expressed through a vector, i.e. PC1. However, this is not a guarantee of efficiently capturing the non-stationarity nature of these data sets. In other words, this explanatory power lives on a linear subspace representation capturing information generated by non-stationary and non-linear data. This argument suggests that such a linear subspace has a good local explanatory power rather than a global one and will be further assessed through the kPCs non-linear solutions of subsection 4.6 and within subsection 4.7.

Appendix F refers to a qualitative analysis, and Appendix G to a quantitative one. Fig. 31 in Appendix F shows three panels. The top panel refers to the PCs of the pollution data, the middle panel refers to the PCs of the climate data and the bottom one to the PCs of the financial data. In the first two plots, we show results of the PCs averaged by counties and by quarters, while, for the financial data set, we observe the first three PCs by county. Fig. 33 in Appendix G shows equivalent panels, i.e. for the pollution, climate and financial data, but this time boxplots of the PCs are shown instead. For the pollution and climate PCs, averages across quarters and counties are considered.

By focusing on the top panel of Fig.31, the standard error of the splines in the last quarter, Q4, is higher than in the other quarters, approximately for every county. Exceptions are Los Angeles, Orange, and Riverside. This may suggest that the variation of pollution variables in the last quarter is generally higher than in the rest of the year, except for the three mentioned counties. Fig. 33 presents the correspondent boxplots for the first three PCs. Particularly the boxplots of PC1 and PC2 show higher variation in Q4, except for Los Angeles, Orange, and Riverside, as expected, confirming the results of Fig. 31. Such analyses, however, do not reveal any details about the strength of the captured variation and how this might differ across counties or how each PC captures this. An index analysis subsection is given below to provide this information (see subsection 4.7). The second panel of Fig. 31 refers to the climate data set. As for the pollution data set, the fitting spline of the PCs in Q4 presents a higher standard deviation than the other quarters suggesting a stronger variability of the underlying data present in that time window. However, in Contra Costa and San Joaquin, an equivalent high standard deviation can be observed in Q1, while, in Los Angeles, in Q2.

Percentage of Variance PCs									
	Pollution			**Climate**			**Financial**		
County	**PC1**	**PC2**	**PC3**	**PC1**	**PC2**	**PC3**	**PC1**	**PC2**	**PC3**
Alameda	**0.560**	0.251	0.143	**0.506**	0.316	0.176	0.121	0.093	0.071
Los Angeles	0.485	0.248	0.184	**0.542**	0.309	0.147	0.195	0.148	0.134
Napa	**0.575**	0.287	0.115	0.469	0.324	0.205	**0.533**	0.466	0.128
San Diego	**0.501**	0.385	0.113	**0.534**	0.311	0.154	0.189	0.108	0.102
San Francisco	**0.570**	0.287	0.112	**0.656**	0.343	0.000	0.220	0.140	0.098
San Joaquin	**0.548**	0.265	0.123	**0.519**	0.307	0.172	0.184	0.128	0.121
San Luis Obispo	**0.573**	0.250	0.141	**0.510**	0.306	0.183	0.296	0.154	0.148
Santa Clara	**0.745**	0.162	0.067	0.483	0.310	0.206	0.126	0.118	0.083
Santa Cruz	**0.557**	0.251	0.150	**0.512**	0.314	0.173	0.178	0.137	0.104

Table 8: Table describing the percentage of variance captured by the first three PCs of each data set, i.e. pollution, climate and financial. Every row shows each considered county in the analysis. Results are between 0 and 1.

Overall, Q3 provides the best fit with the least standard variation observed, suggesting lower variations. By looking at Fig. 33, the second panel shows boxplots of the climate data set. Compared to the qualitative analysis, assessing

how the variations changes across the different quarters is harder in practice. However, the most minor variability is confirmed in Q3, where the boxplots show more minor variations in all the PCs.

The last panels of both Fig. 31 and Fig. 33 present the results for the financial PCs. Even though the spline fitting of Alameda and Napa show lower standard deviations (for Napa, this is justified for the low number of green bond issues, i.e. only three, and hence three samples), no significant differences are observed across the counties. A remark is that in Fig. 33, the PCs boxplots referring to Napa appear quite spread out; nevertheless, Napa has only three issued bonds, and only three points contribute to the shape of the boxplots.

4.6 Non-Linear Solution: kPCA

This section presents an equivalent analysis to the one performed above, but focuses on the kPCs. Table 9 shows the percentage of variance of the kPCS. Compared to the PCs, the variability of the underlying data sets is split as 40-50% for PC1, 20-30% for PC2 and 10-20% for PC3. This is constant across the three data sets. Furthermore, in the case of the financial data and, compared to the PCs, the kPCs consistently detect similar percentages. While PCA shows a good local linear approximation with PC1 and a quick decay of the other bases implying a low global explanatory power, overall the kPCs carry a much more uniform explanatory power, nearly equally distributed across the three kPCs. Given the high non-stationarity levels of the data, the kPCs, which represent functions, are expected to perform better than the PCs, being vectors instead. This uniform distribution of the explanatory powers of kPCs in terms of carried percentage of variance will be further discussed in Subsection 4.7.

Fig. 32 in Appendix F is structured similarly to Fig. 31, thus presenting three panels, one for each data set, i.e. pollution, climate, and financial ones. The first two panels present the kPCs by county and quarters. The third panel refers to the KPCs for the financial data set and the counties considered. As mentioned, we fitted a penalised spline and provided the standard error. Fig. 34 in Appendix G shows the boxplots referring to the correspondent kPCs.

By focusing on the first panel of Fig. 32 that refers to the kPCs of the pollution data set, it is possible to observe how, compared to the case of PCs, there is not a specific quarter for which the spline fitting presents higher levels of standard deviation. However, counties such as Contra Costa, Kern, San Joaquin, San Mateo, and Santa Clara show fitted splines with larger standard deviation estimates. If one considers the correspondent pollution boxplots given in Fig. 34, it is again possible to observe that compared to PCs, no specific quarter shows a higher level of variability in terms of PCs. However, several counties like Alameda, Contra Costa, San Joaquin, San Mateo and Solano show more spread boxplots, particularly for kPC1 across all the quarters, than the rest. Regarding the case of the climate kPCs, the second panel of Fig. 32 shows low levels of fitted spline standard deviations overall, except for the fourth quarter (for most counties). Counties behave similarly in this respect, apart from San Joaquin in Q1, where standard errors appear way higher. Fig. 34 presents the correspondent boxplots of the kPCs of the climate data. It is hard in practice to assess any particular variation, even though it seems that kPC2 varies more across counties and quarters.

Percentage of Variance kPCs									
	Pollution			**Climate**			**Financial**		
County	**kPC1**	**kPC2**	**kPC3**	**kPC1**	**kPC2**	**kPC3**	**kPC1**	**kPC2**	**kPC3**
Alameda	0.466	0.293	0.241	0.423	0.366	0.210	0.453	0.319	0.229
Los Angeles	0.390	0.334	0.276	0.497	0.349	0.153	0.429	0.357	0.214
Napa	0.422	0.329	0.249	0.501	0.329	0.170	0.500	0.500	0.000
San Diego	0.495	0.318	0.187	0.409	0.366	0.225	0.595	0.244	0.162
San Francisco	0.430	0.322	0.249	0.403	0.322	0.275	0.406	0.315	0.279
San Joaquin	0.452	0.330	0.218	0.510	0.338	0.152	0.468	0.270	0.263
San Luis Obispo	0.464	0.330	0.206	0.432	0.359	0.210	0.587	0.277	0.136
Santa Clara	0.478	0.283	0.239	0.412	0.396	0.193	0.402	0.342	0.256
Santa Cruz	0.498	0.311	0.192	0.513	0.341	0.147	0.340	0.330	0.330

Table 9: Table describing the percentage of variance captured by the first three kPCs of each data set, i.e. pollution, climate and financial. Every row shows each considered county in the analysis. Results are between 0 and 1.

In the case of the financial data set given in the last panel of Fig. 32, overall, a similar behaviour appears, apart from Napa, in which the standard deviation of the fitted spline is zero due to the small sample size of issued green bonds, i.e.

the number of samples. The last panel of Fig. 34 refers to the kPCs boxplots. Compared to the results obtained for the PCs, which behave similarly across all the counties (given in Fig. 32, last panel), the kPCs appear to vary significantly. kPC1 varies the most except for San Francisco and Santa Cruz. kPC2 shows fewer spread boxplots, again in Santa Cruz except for Napa, where only three data points are employed. kPC3 also shows fewer spread boxplots, particularly in Los Angeles, Napa, San Diego and Santa Cruz. As for the PCs, an index analysis is now given in the following subsection.

4.7 Index Analysis

To compare the results obtained from the different PCs and kPCs and their performances in capturing the structure of the underlying data sets as well as their explanatory power presented above, we employ what is referred to in machine learning as "Empirical Centered Kernel Alignment" (CKTA) [49]. Such a technique is employed within kernel methods to identify optimal kernel hyperparameters by computing a distance between the empirical covariance matrix of the original data and the kernel Gram Matrix computed with the fitted hyperparameters of interest through a certain measure of kernel similarity or alignment. Several alignment computation methods have been proposed. The idea of [49] considers two Gram Matrices $\mathbf{K}_1$ and $\mathbf{K}_2$ and define the CKTA as

$$\hat{\rho}\left(\mathbf{K}_1, \mathbf{K}_2\right) = \frac{\langle \mathbf{K}_1^c, \mathbf{K}_2^c \rangle_{\mathcal{H}}}{\sqrt{\langle \mathbf{K}_1^c, \mathbf{K}_1^c \rangle_{\mathcal{H}} \langle \mathbf{K}_2^c, \mathbf{K}_2^c \rangle_{\mathcal{H}}}} \quad \in \quad [-1, 1] \tag{44}$$

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 $\mathbf{K}$. 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

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

From a statistical point of view, the alignment could be considered an equivalent of measure to the Pearson coefficient of correlation between two random variables, i.e. $\mathbf{K}_1(\mathbf{x}, \mathbf{x}')$ and $\mathbf{K}_2(\mathbf{x}, \mathbf{x}')$, generated by an equally weighted uniform distribution over pairs $(\mathbf{x}, \mathbf{x}')$. The idea is to interpret the CKTA as the cosine distance between two-dimensional vectors. We employ CKTA to compute the distance between the empirical covariance matrix and the Gram Matrices computed with PC1, PC1 and PC2, PC1 and PC2, and PC3 first. Afterwards, an equivalent procedure is performed with kPC1, kPC1 and kPC2, kPC1 and kPC2 and kPC3. This is done at a county level for every considered data set; for example, by computing the distances between the empirical covariance matrix of the engineered pollution features of Alameda and the different PCS extracted on such features (as introduced just yet) and the different kPCs. In this way, we can assess which bases better capture the structural variability of the proposed features. Using both PCs and kPC, we can evaluate whether this variability is linear and/or stationary or not.

A high explanatory power obtained through the above captured variance percentage does not necessarily imply a high cKTA. Indeed, PCs detect local properties of the given data rather than global structures; hence will tend to show lower cKTAs even though PC1 has very high variance percentages and, therefore, high explanatory power. Each considered matrices employed in the cKTA computations present an approximation matrix of the empirical covariance matrix. A way to measure the quality of this approximation is by using cKTA. Therefore, cKTA could be interpreted as a model ranking procedure.

Results of the CKTA are given in Tables 10 and 11 for pollution and climate data sets, respectively. Note that in these tables, we report results only for the counties used in the constructed CCA models. For completeness, Tables 24 and 25 in Appendix H show all the counties considered in each data set. For each table, we present in columns the different combinations of PCs and kPCs studied; in rows, instead, we provide the counties since the PCs and the kPCs extraction has been done at this level. We highlighted alignments that are superior to 70% as, in practice, a 70% level of alignment represents high variability captured. By focusing on Table 10, it is possible to observe how the PCs, whichever is the selected combination, do not show a CKTA higher than 70% in any county apart from Santa Clara. In the case of the kPCs, instead, the combinations of first and second as well as first, second, and third kPCs show alignments of 70% and 80% in several counties. Overall, the performances increase with the number of bases included in the CKTA computations, i.e. the best performances are obtained when all the three kPCs are retained. Note that counties such as Alameda, Los Angeles, San Diego and Santa Clara are the most populated counties, with a population ranging between 1 to 3 million. In these cases, the relative difference between kPC1 and kPC1 and kPC2 appears to be more prominent than other counties with fewer populations, hence suggesting that kPC2 tends to

detect significant variability within more populated counties. Table 24 of Appendix H show the same reasoning for Contra Costa, Orange, San Bernardino. Overall, the results indicate that kPCs better capture the underlying engineered pollution features than PCs.

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**

Table 10: 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.

Table 11 refers to the case of climate, thus showing CKTA between the engineered climate features and the extracted PCs and kPCs, respectively. Compared to the pollution data sets case, this time, PCs present higher CKTAs within different counties, achieving results superior to an alignment of 80% within Marin, Napa, Riverside, San Mateo, Solano, and Yolo. In the case of kPCs, several counties have been efficiently captured by using kPC1 and kPC2, and kPC1, kPC2, and kPC3. The latter is indeed the one that provides the best overall performance, with several counties above 90% of alignment, suggesting that the first three kPCs efficiently capture the variability of the proposed climate features. Consistently with the pollution case, counties with major populations show relative differences between kPC1 only and kPC1 and kPC2 or kPC1, kPC2 and kPC3, which are larger than counties with lower population levels. Therefore, in terms of the pollution case, the second kPC and the third kPC appear to contribute much more in the case of more populated counties. In the remaining counties, i.e., counties with a lower population, Table 25 in Appendix H shows that the first kPCS appear to perform well in the case of Madera, Riverside, and Sacramento. In the other counties, the first kPC alone does not show an alignment higher than 70%. The best performances, as in the pollution data set, are provided by the three kPCs together, achieving an alignment of 80% or higher in most cases. High alignments are also achieved in the case of PCs, suggesting that the engineered climate features carry a more linear variability than the pollution ones.

Lastly, Table 12 refers to the CKTA results for the financial data set. 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. Regarding the kPCs, higher levels of alignments are identified for all counties except for San Francisco and Santa Clara. In the cases of Alameda, Los Angeles, and Santa Cruz, while kPC1 or kPC1 and kPC2 do not appear to provide good alignments, when all three kPCs are employed, CKTAs higher than 70% are instead found. For Napa, San Diego, San Joaquin, and San Luis Obispo, every tested combination of kPCs well perform. As foreseen, the high complexity of this data cannot rely on linear PCs. This means that the data cleaning procedure accompanied by the hot encoding and the usage of the Jaccard kernel for the treatment of categorical variables strongly promotes the kPCA method, given the high performances achieved.

As presented, the explanatory power, expressed in terms of captured percentage of variance, of the PCs and the kPCs differ. Indeed, in the linear case, PC1 carries the great majority of the underlying variability, while in the non-linear case, this is distributed uniformly across the first three bases. Several points must be raised at this stage. Firstly, PCs are linear bases and represent vectors. On the contrary, the kPCs are non-linear functions computed in the kernel space. Even if the linear projection of PC1 captures a great deal of variability, it does not efficiently approximate the non-stationary nature of the underlying data. The results of the obtained cKTAs in this section can further support such

an argument. On the contrary, while the kPCs carry a more uniform variation of the original data, these are evaluated functions in the kernel space dealing with non-stationarity, providing a way better approximation of the data. This is supported by the cKTA results provided in this subsection.

To further understand cKTA and how this works in practice, Figure 17 presents an example for the pollution data set and the county of Alameda. Panel (a) represents the heatmap of the covariance matrix of the engineered features for the pollution data averaged by quarters. Hence, instead of considering a time series of 10 years, this data summarises information by quarters to make variations more visible. If one then focuses on panels (b), (c), and (d), these represent the covariance matrices computed with PC1 as column vector, PC1 and PC2 as column vectors and PC1, PC2 and PC3 as column vectors, respectively. Panels (e), (f) and (g) show equivalent matrices, but, this time, the kPCS have been equivalently employed.

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 11: 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.

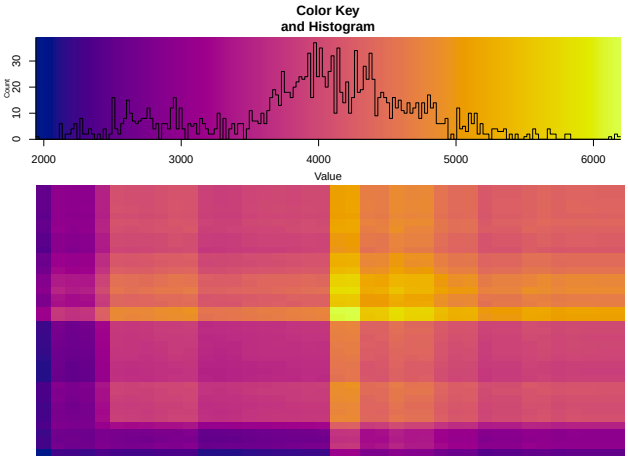
Results of Centered Kernel Target Alignment - Financial Dataset

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**

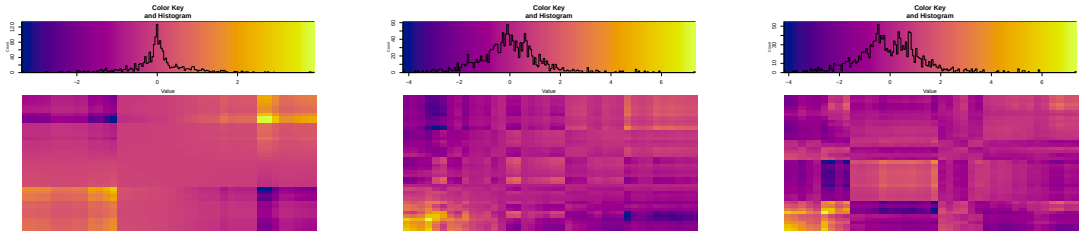
Table 12: cKTA results for the financial data set. Each row shows a considered county, while, in the columns, we have the different approximation matrices used for the cKTAs. The first column presents the cKTAs calculated using the covariance matrices of the financial data and the rank-one approximation 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.

The cKTAs in Tables 10 for the county of Alameda are computed as the alignments, following Equation (44), between the matrix of panel (a) and each other of the matrices in the other panels. Note that this specific example refers to the

pollution data, averaged over quarters for visualisation reasons, but a similar approach has been performed on the ten years time series. A further description note is that these matrices must be centred for the cKTA to be computed. As before, this operation has been omitted for visualisation purposes since hiding the relevant sought information. This procedure for computing the cKTAs has been applied to all the other counties and data sets.



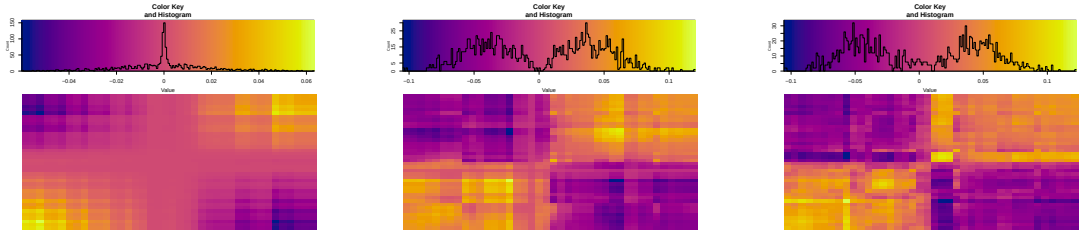
(a) Original engineered pollution data features averaged by quarters.



(b) PC1.

(c) PC1 and PC2.

(d) PC1, PC2 and PC3.



(e) kPC1.

(f) kPC1 and kPC2.

(g) kPC1, kPC2 and kPC3.

Figure 17: Figure supporting the cKTA computations. Panel (a) represents the heatmap of the covariance matrix of the original engineered features of pollution data averaged by quarters for the county of Alameda. The first row refers to the PCs, and the second row refers to the kPCs. Both bases have also been averaged by quarters. Hence, instead of considering a time series of 10 years, this data summarises information by quarters to make variations more visible. By focusing on the second row (plots (b), (c) and (d)), panel (b) represents the heatmap of the rank-one approximation covariance matrix obtained using the column vector of PC1. Panel (c) represents the rank-two approximation covariance matrix obtained using the two-column vectors PC1 and PC2. Panel (d) represents the rank-three approximation covariance matrix obtained using the three-column vectors PC1, PC2 and PC3. Equivalent reasoning applies to the third row (plots (e), (f) and (g)) where, this time, kPC1, kPC1 and kPC2, kPC1 and kPC2 and kPC3 have been used instead to produce rank-one, rank-two and rank-three approximation covariance matrices, respectively. Note that when the CKTA is computed, each matrix is centred to get values in the same ranges. No matrix has been scaled for visualisation purposes and to better observe how the kPCs outperform the PCs.

By focusing on the heatmaps of the PCs and the kPCs and moving from left to right, it is possible to observe how the components capture different structures of the original data. Indeed, these matrices represent different approximation ranks of the original data and, by considering bigger ranks, i.e. rank-one, rank-two and rank-three, the given

approximation, measured with the cKTA alignment, tends to increase. Furthermore, the kPCs capture much more non-stationary and non-linear behaviour, suggesting better kPCA-CCA performances.

4.8 The Cross-Correlation with CCA

This section is dedicated to analysing the PCA-CCA and kPCA-CCA results. We first observe the correlation matrices of the PCs and the kPCs considered to assess the within-relation and between-relation of these bases describing the data sets introduced above. Afterwards, an overall assessment of the proposed model is provided. We will examine statistical tests to understand the results of the canonical correlation for PCA and kPCA. Subsequently, for results that show statistically significant results, we will assess the redundancy index. Lastly, we will look at the canonical correlation by considering the structure coefficients or canonical loadings. 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 the obtained results.

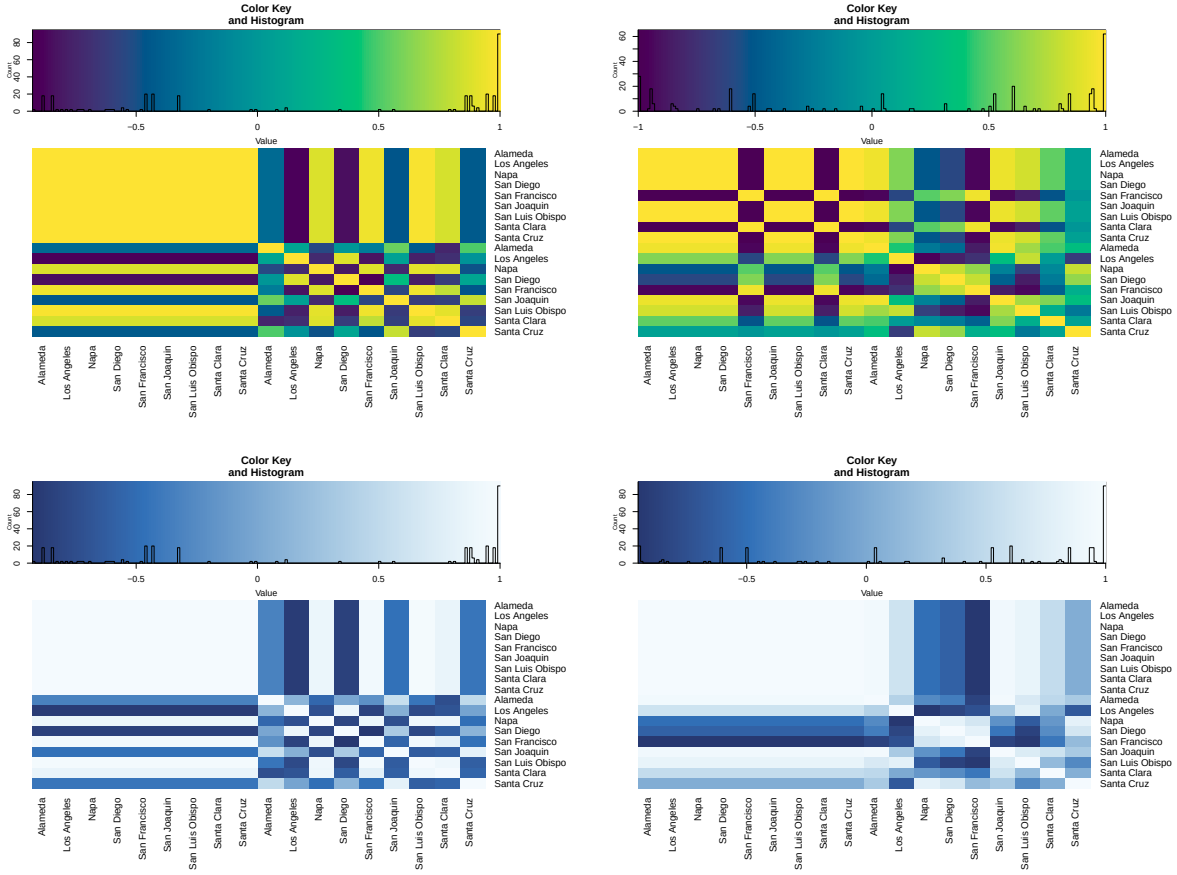


Figure 18: Heatmaps of correlation matrices between financial and pollution PCs (top panels) and financial and climate PCs (bottom panels). The left panels refer to the PC1 correlation matrices and the right panels to the PC2 correlation matrices.

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.

Before moving on to the output of the CCA, we first observe the correlation matrices of the PCs and the kPCs extracted on the pollution and financial data sets and the climate and financial data sets, respectively. Such matrices are represented in the heatmaps of Fig. 18 and Fig. 19. The top panels of Fig. 18 refer to the heatmaps of the correlation of PC1 (left) and PC2 (right) extracted by the financial and pollution data, respectively. The bottom panels represent the heatmaps of the correlation of PC1 (left) and PC2 (right) extracted by the financial and climate data. An equivalent structure is followed in Fig. 19 for the kPCs, where the top panels are the heatmaps of the correlation of kPC1 and kPC2 extracted by the pollution and financial data and the bottom panels represent the heatmaps of the correlation of kPC1 and kPC2 extracted by the climate and financial data sets. Each matrix is structured by having the two blocks of the main diagonal with the within correlation matrices, i.e. the correlation of, for example, PC1 of one data set (financial) for the considered counties and in the second block, the correlation of PC1 for the considered counties of the second data set (pollution/climate). Instead, in the off-diagonal blocks, there will be the between or cross-correlation between PC1 of one data set (pollution/climate) and PC1 of the other (financial) data set in one block and its correspondent transpose in the other.

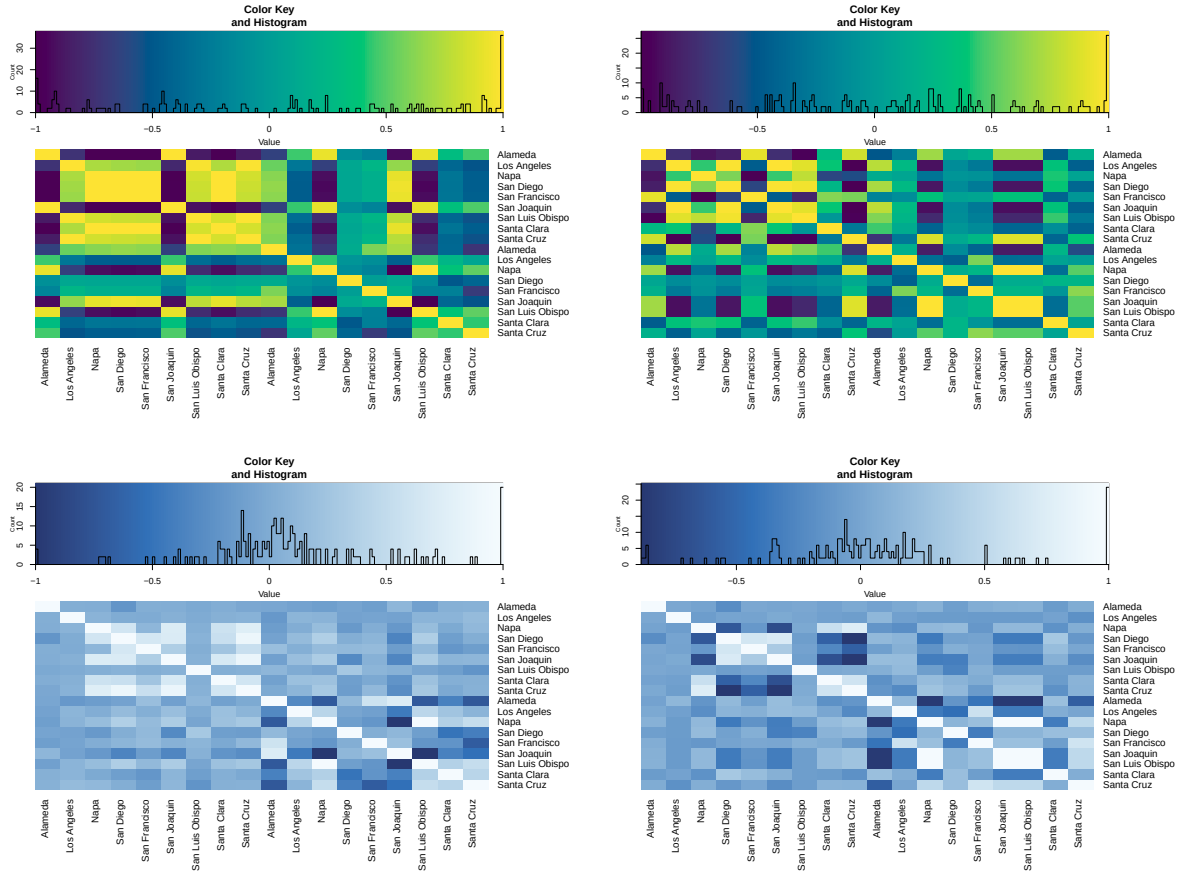


Figure 19: Heatmaps of correlation matrices between financial and pollution kPCs (top panels) and financial and climate kPCs (bottom panels). The left panels refer to the kPC1 correlation matrices and the right panels to the kPC2 cross-correlation matrices.

The results of the heatmaps show how the PCs display a way different correlation structure compared to the kPCs. Indeed, given the similar structure of the PC1 pollution/financial matrix and PC1 of the climate/financial one, the financial variables seem to drive the correlation results. The top-left blocks appear to carry maximum correlation across all the counties in both matrices. This is mainly due to the presence of many categorical variables within this data which are strongly leading the variation captured by PC1. A Similar reasoning can be seen for the matrices of PC2, with an exception for the heatmap of PC2 of pollution/financial where the correlation structure appears to be more

variable. If one focuses on heatmaps of the kPCs, it is possible to observe how different and variate is the correlation between these bases. This indeed suggest that the kPCs detect variation of the underlying data more efficiently. This work's main objective is to consider the cross-correlation matrices of these heatmaps and perform CCA on it so that the within correlation of these data will be removed. However, if the CCA was directly applied on the raw data, the high levels of non-stationarity and non-linearity contained within the data would have polluted the cross-correlation structures and affecting the final results which, indeed, would have been unreliable. A further point to make at this point is that, by using kPCA, the categorical variables of the financial data can be efficiently encoded and treated with the use of the Jaccard kernel. In such a way, the non-stationarity structure of the raw data can be maintained, as shown in the within correlation blocks of the heatmaps of the kPCs, where, compared to the PCs ones, there is no evidence of such high correlation due to the presence of such categorical variables. If one consider, for example, the correlation matrix of kPC1 for pollution and financial data, the top-left block of this matrix is the within correlation block of kPC1 computed for the financial data. It is possible to observe how this correlation appears to be non-stationary and different from the correspondent block of the heatmaps of PC1 in Fig. 18, top left panel. Therefore, the use of kPCA does not only deal with non-stationary and non-linear data but also facilitates the use of a different type of variables through embedding different kernels, i.e., radial basis function for numerical and Jaccard for categorical.

Canonical Correlation Summary Financial Data Set vs Pollution Data Set

PC1						
Main Results			**F Test for Canonical Correlations (Rao's F Approximation)**			
CV	ρ^*	${\rho_i^*}^2$	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

Table 13: PC1-CCA Model Assesment.

PC2						
Main Results			**F Test for Canonical Correlations (Rao's F Approximation)**			
CV	ρ^*	${\rho_i^*}^2$	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

Table 14: PC2-CCA Model Assesment.

Canonical Correlation Summary Financial Data Set vs Climate Data Set

PC1						
Main Results			**F Test for Canonical Correlations (Rao's F Approximation)**			
CV	ρ^*	${\rho_i^*}^2$	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

Table 15: PC1-CCA Model Assesment.

PC2						
Main Results			**F Test for Canonical Correlations (Rao's F Approximation)**			
CV	ρ^*	F	${\rho_i^*}^2$	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

Table 16: PC2-CCA Model Assesment.

Afterwards, we observe the results for assessing the PCA-CCA and kPCA-CCA models overall. These results are provided in Tables 13 to 16 for the PCA-CCA, while in Tables 17 and 20 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 13 to 16, 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. No canonical correlation is high enough and significant for any selected PCs according to the F test. This strongly supports our explanation for the correlation matrices and that, given that the underlying data sets carry such a strong non-stationary variation, the PCs do not capture any of this global non-stationarity present in the data. PC1 seems to have a strong explanatory power based on the percentage of captured variance given in Subsection 4.5. However, as presented in

subsection 4.7, the alignments provided with the original empirical covariance matrices of such basis are low almost in every county and every data set. This suggests that if the interest is analysing cross-correlation, one should use non-linear and non-stationary bases functions ad the kPCs. Given such results for the PCA-CCA, we will not show any other part of the CCA analysis of this benchmark since it will be insignificant and hardly interpretable.

The results for kPCs are instead provided in Tables 17 to 20. Tables 17 and 18 show the results for kPC1 and kPC2, respectively, of the pollution versus the financial data sets. It is possible to observe how, 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 18, an equivalent reasoning to kPC1 can be applied. Indeed, the canonical correlations of the first two canonical functions correspond to 1.000 and 0.881, respectively, with squared canonical correlations of 0.999 and 0.776, hence showing high levels of variance shared by both the canonical variates of each canonical function. Again, all the canonical functions except for the nineth one are significant according to the F test. However, only the first two will be retained and considered in the analysis.

Taking into account now Tables 19 and 20, one can then analyse the results of the CCA for kPC1 and kPC2 of the climate/financial data. This time, across both tables, the only canonical correlation higher than 0.700 is one of the first canonical functions of kPC1 with a canonical correlation of 0.815 and a squared canonical correlation of 0.712. All the others, for both kPC1 and kPC2, are below such a threshold, hence suggesting a low level of correlations between these two modes of variations extracted on financial data and climate data. Such a result is expected since, in practice, a bigger time frame must be studied for a green bond issued to affect or, as sought in this work, to be correlated with climate variables. Furthermore, by focusing on the F test results, less canonical functions appear to be significant compared to the pollution 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.

Overall, 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.

The following step of the analysis focuses on the redundancy index, or redundancy plots for the case of kPC1 and kPC2, referring to the CCA analysis of pollution and financial data. These are given in Fig. 20, where the left plot refers to kPC1 and the right one refers to kPC2. Note that an equivalent analysis was carried out on the case of financial/climate, but the results were not significant and, therefore, not included for space reasons. Remark that 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. In particular, we provide in Fig. 20 two plots, one for kPC1 and one for kPC2, where we can observe how the total fraction of variance of the first kPC of all the counties extracted on the financial variables is accounted by the first kPC of all the counties extracted on the pollution variables, through each canonical variate, and, vice versa. Hence, in the left plot, X in the name of the x-axis refers to the first kPCs of the financial variables and Y in the name of the x-axis to the first kPCs of the pollution variables. The y-axis represents the fractional variance explained. In the right plot, the same reasoning applies but for kPC2. Note that the redundancy is given in total, i.e. the red bar, and for individual canonical variates, i.e. the remaining bars, as presented in the legends. It is possible to observe how, for both kPC1 and kPC2, the total variance of the financial kPCS accounted for by the correspondent pollution kPCs is approximately 60%; while the total variance of the pollution kPCs accounted for by the correspondent financial kPCs is approximately 80%. Hence, while these redundancy indices are high overall, interestingly, the variance of the pollution kPCs captured by the financial appears to be higher. Since high redundancy suggests an increased ability to predict, this indicates that the financial kPCs tend to explain the pollution kPCs efficiently. Particularly, since the kPCs represent modes of variations, the financial variables should, in an ideal world, be predictive of the pollution in the sense that, in a very mere way, a higher level of financial green bonds must be able to predict their pollution effect. Such a plot strongly supports this idea. However, note that in this prediction, there is no strength of the direction, meaning it is not clear if the effect of the financial kPCs is positive or negative concerning pollution kPCs. This sustains that the financial variables must be predictive of pollution. Further, in the case of kPC1, the first canonical variate appears to have the majority of variance explained, while in the second kPC, this variance is lower. This further reinforces this analysis since kPC1 is the leading basis function carrying the greatest variations in kPCs.

Canonical Correlation Summary Financial Data Set vs Pollution Data Set

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

Table 17: kPC1-CCA Model Assessment.

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

Table 18: 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^*	${\rho_i^*}^2$	F	df_2	df_1	Pr(>X)
1	**0.815**	**0.712**	27.601	81	6355.2	$\mathbf{< 2.2e-16}$
2	0.673	0.453	21.085	64	5676.3	$\mathbf{< 2.2e-16}$
3	0.491	0.241	13.377	49	5000.0	$\mathbf{< 2.2e-16}$
4	0.423	0.179	9.939	36	4328.2	$\mathbf{< 2.2e-16}$
5	0.259	0.067	6.074	25	3664.3	$\mathbf{< 2.2e-16}$
6	0.215	0.046	5.067	16	3016.0	$\mathbf{1.53e-10}$
7	0.165	0.027	3.686	9	2404.7	$\mathbf{0.001}$
8	0.073	0.005	1.397	4	1978.0	0.232
9	0.014	0.000	0.202	1	990.0	0.652

Table 19: kPC1-CCA Model Assessment.

kPC2						
Main Results			**F Test for Canonical Correlations (Rao's F Approximation)**			
CV	ρ^*	F	${\rho_i^*}^2$	df_2	df_1	Pr(>X)
1	0.664	0.441	22.827	81	6355.2	$\mathbf{< 2.2e-16}$
2	0.572	0.327	18.038	64	5676.3	$\mathbf{< 2.2e-16}$
3	0.515	0.265	14.406	49	5000.0	$\mathbf{< 2.2e-16}$
4	0.374	0.140	10.309	36	4328.2	$\mathbf{< 2.2e-16}$
5	0.347	0.120	8.523	25	3664.3	$\mathbf{< 2.2e-16}$
6	0.242	0.058	5.093	16	3016.0	$\mathbf{1.294e-10}$
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

Table 20: kPC2-CCA Model Assessment.

The following step of the analysis looks at the kPCA-CCA results in more detail 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 of financial/pollution data sets and financial/climate data sets. These results are given in Tables 21 and 22, where the left tables refer to pollution/financial while the right tables refer to climate/financial. 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. 21 and Fig. 22, for pollution/financial and climate/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. Remark that a structured coefficient represents the equivalent 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 synthetic 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. Hence, 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. In such a way, 70% is always the considered threshold and quite significant. 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.

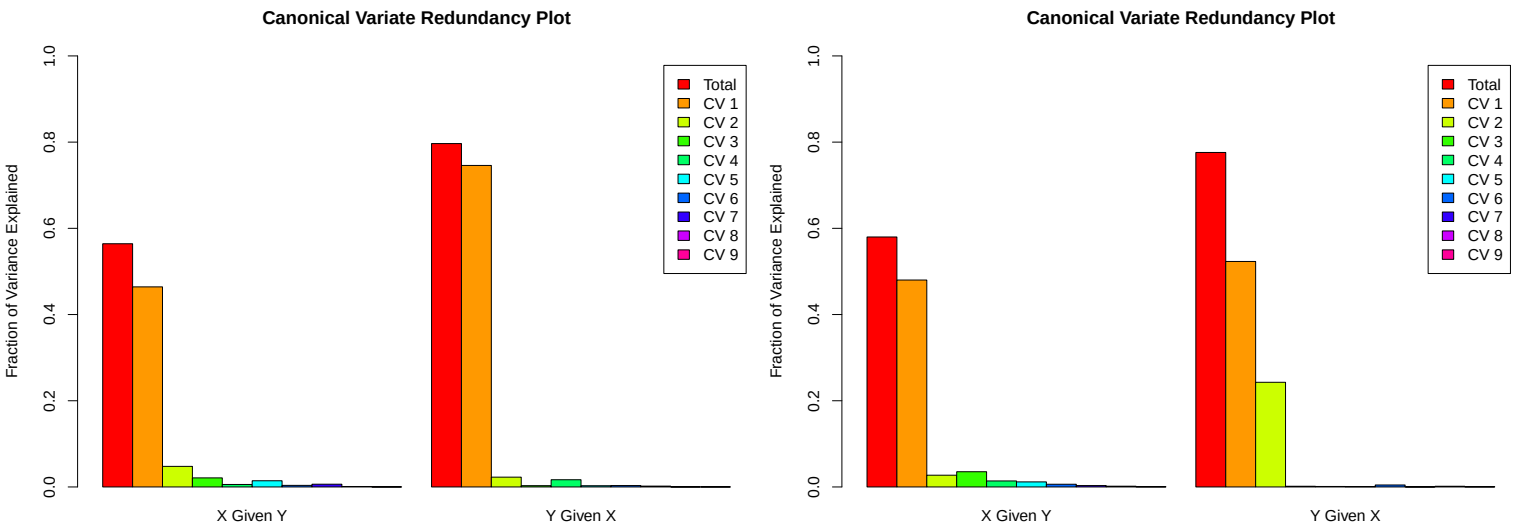


Figure 20: Redundancy Plots. The redundancy index determines how much of the variance is accounted for in one set of variables by the other set of variables. These plots consider the case of kPCs extracted from the financial and pollution data, and the left plot refers to kPC1, while the right plot refers to kPC2. In both sets of plots, the X in the name of the x-axis refers to the kPC of the financial variables and Y in the name of the x-axis to the kPC of the pollution variables. The y-axis represents the fractional variance explained. The plots show these quantities in total, i.e. the red bar, and by canonical variates, i.e. the remaining bars, as given in the legends.

If one focuses on the top panel of Table 21, 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; therefore, we analyse only this one. Following the above reasoning, it is possible to observe how the financial structure coefficient of kPC1 of Alameda and San Joaquin are positively correlated and how the financial structure coefficients of kPC1 of Napa and San Luis Obispo are positively correlated. Therefore, these two groups are negatively correlated, given the opposite signs of their structure coefficients. By looking at the pollution kPC1, the structure coefficients of kPC1 of Alameda and San Joaquin are positively correlated. In contrast, the structure coefficients of kPC1 of Los Angeles, Napa, San Diego, San Francisco, San Luis Obispo, Santa Clara and Santa Cruz are positively correlated. These two groups are negatively correlated, given the opposite signs of their structured coefficients. Interestingly, the structure coefficients of kPC1 of Alameda and San Joaquin of the financial data are negatively correlated with their corresponding pollution ones. A similar reasoning for the structure coefficients of kPC1 of Napa and San Luis Obispo can be made, suggesting a negative correlation between the first kPC of pollution and financial. Hence, this indicates that these counties' greatest mode of variation for the financial variables drives in the opposite direction to its corresponding pollution one in terms of correlation. As for the redundancy index discussion, what is unclear in this scenario is if the correlation implies a positive or negative impact. A further point is the negative correlation between the financial kPC1 of Alameda and Sa Joaquin with Napa and San Luis Obispo. This could suggest that given that Alameda has issued the greatest number of bonds in the green financial market, this is negatively correlated with the other counties' issuances. This reasoning will be better explained when kPC2 is analysed. For the signs of the structured coefficients for the pollution kPC1 concerning the different counties, in this case, an interpretation is more complicated. The fact that the pollution structured coefficients of kPC1 of Alameda and San Joaquin are negatively correlated to the ones of Los Angeles, Napa, San Diego, San Francisco, San Luis Obispo, Santa Clara, and Santa Cruz is certainly more challenging to understand and/or motivate. This sign direction could be used if jointly analysed with the financial structured coefficients. Alameda and San Joaquin have positive financial structure coefficients and negative pollution structure coefficients. This might indicate that, in practice, green bonds have a positive effect in these counties. For Napa and San Luis Obispo, two different roads must be taken in analysing. Napa has only three green bonds issued in the market (at least at the moment of the analysed data); therefore, the small sample size could have included multicollinearity influencing the sign of the correlation. There is an undoubted correlation, but the direction the sign gives is more uncertain in this analysis. For San Luis Obispo, a negative financial structured coefficient and a positive pollution coefficient could instead imply a negative impact of the issued green bonds. Coming back to other significant pollution structured coefficients with a positive sign, i.e.

Los Angeles, San Diego, San Francisco, Santa Clara and Santa Cruz, the same reasoning applied to San Luis Obispo could also be fair in this case, i.e. for Los Angeles, Santa Clara and Santa Cruz a negative relationship between the structured coefficients of pollution and financial can be found, but the low levels of correlations in the financial kPC1 make this statements not reliable. Significant correlations for these counties will be found in the case of kPC2, and better motivations will be provided. A relevant point is the number of issued green bonds. Fig. 13 shows the number 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.

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**

Table 21: Canonical Correlation Analysis

KPCA-CCA - Financial Data Set vs Climate Data Set

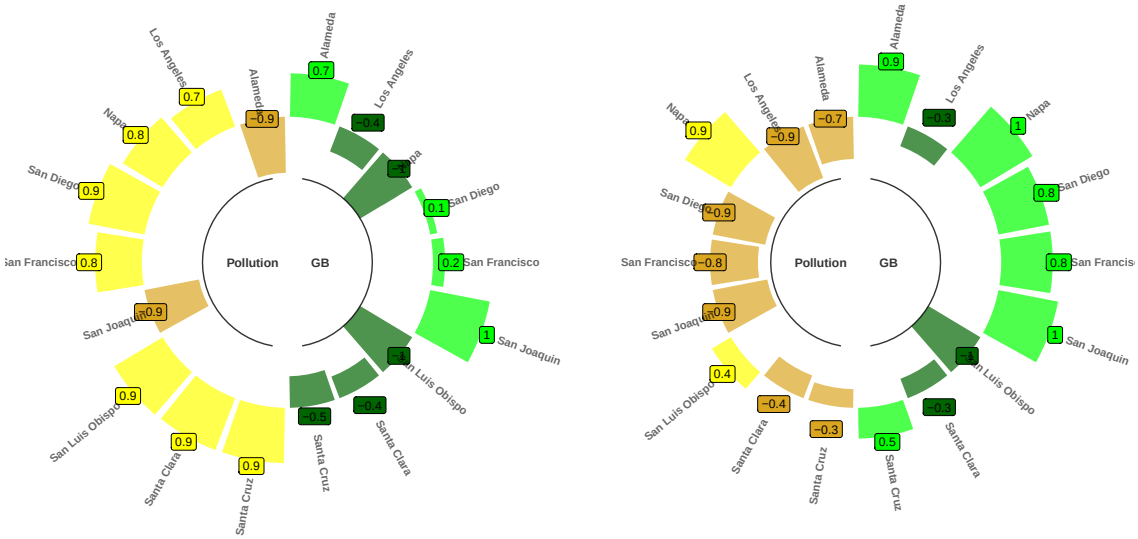
			Structured Coef.		**Squared Structured Coef.**	
Data set	kPC	County	CV 1	CV 2	CV 1	CV 2
Financial	1	Alameda	-0.081	-0.552	0.007	0.305
Financial	1	Los Angeles	0.307	0.368	0.094	0.135
Financial	1	Napa	0.556	0.310	0.309	0.096
Financial	1	San Diego	-0.135	-0.558	0.018	0.311
Financial	1	San Francisco	0.130	-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

Table 22: Canonical Correlation Analysis

We now move to the analysis of CCA on kPC2 for pollution and financial data sets. If one follows the same reasoning conducted for kPC1, then reinforcing results can be found. The positive financial structured coefficient of Alameda and its negative corresponding pollution one is found again. In this kPC, an equivalent relationship is found for the states of Napa (contrasting the findings of kPC1 but reinforcing the problem of the small sample size for the financial variables), San Diego, San Francisco, San Joaquin (as for kPC1). Contrary to what found for kPC1 for some of these counties, there is an equivalent relationship to the one of Alameda, hence positive financial structured coefficients and negative pollution structure coefficients, suggesting the positive impact of the green bonds for these counties. For San Luis Obispo there is a significant financial structure coefficient with the same direction of kPC1 but a not significant one for the pollution. However, the structure coefficients of the second canonical variate appear to be significantly high with the sign in the opposite direction, suggesting a negative impact in this county again, even though this reasoning cannot be formally applied across variates. What is interesting is that while kPC1 appears to capture a multivariate relationship only across Alameda, Napa, San Joaquin and San Luis Obispo, the second kPC instead presents a relationship also across other counties as San Diego and San Francisco, suggesting that these two modes of variations capture different underlying information. This can be due to several factors, particularly the fact that Alameda is probably predominant in the first kPC, making it more difficult to capture the pattern of counties

where high amounts of bonds are issued. Appendix I shows further plots of structured coefficients for more canonical variates (the second one and the third one).

Table 22 shows results for the structured coefficients and squared structured coefficients of the first and second canonical variates of kPC1 and kPC2 for studying climate/financial. It is straightforward to understand that the results for the climate do not show high correlations with very few exceptions. Still, the authors believe this is a relevant result since it requires major attention from the community due to the claim that climate change is highly linked to green bonds. Further research is required in this direction, with larger time spans and more refined advanced cross-correlation methods.



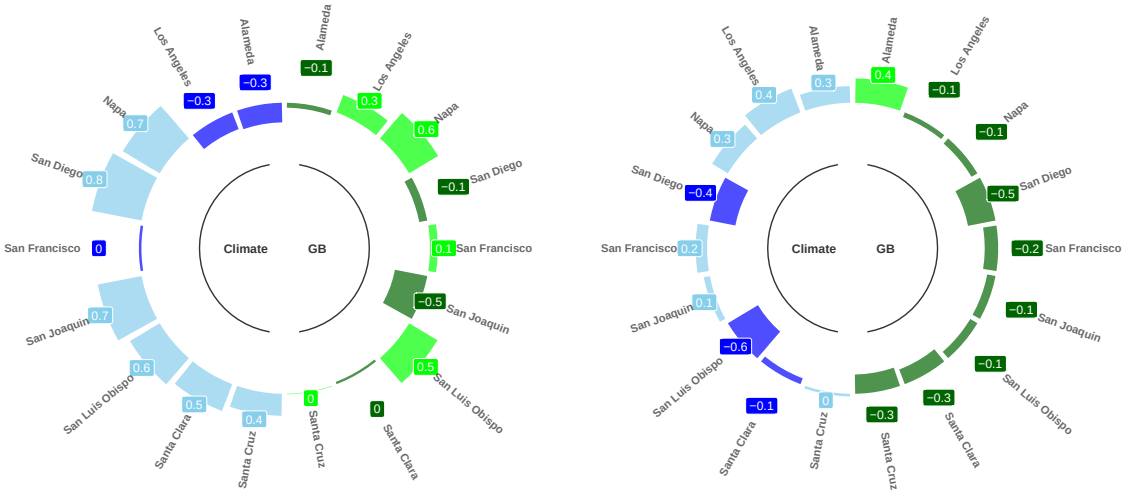
(a) kPC1-CV1.

(b) kPC2-CV1.

Figure 21: Structured coefficients of financial/pollution kPC1 (left) and kPC2 (right) for the first canonical variate. These are presented in Table 21, in the fourth column.

The final part of this section discusses the financial variables' contribution to the obtained result. After the general model assessment, a standard multivariate linear technique would require at this stage to understand how and at which point the original financial variables contribute to the variation in the data, yielding to the final guidelines provided for decision-making processes. 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 variable, one by one, to observe at which level these are captured and, for the numerical variables, we use 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. Each of these two figures displays these measures by variable and by kPC. Hence it shows how each kPC captured one specific numerical or categorical variable. The county presents the results since the county has done the analysis.

By observing Fig. 23, it is possible to observe how the best cKTAs are obtained for the county of San Diego, Santa Clara, and Santa Cruz. In San Diego, the best represented variable corresponds to Muny Source for both kPCs, while, in Santa Clara the best categorical variable is Muny Issue Type. For Santa Cruz, instead, more variables are captured, i.e. Muny Source, Muny Offering Type, and Issuer Industry. For Alameda, it seems that kPC1 is better capturing the categorical variables, while, for Los Angeles and San Francisco is instead kPC2. However, the level of achieved alignments is around 0.4, thus suggesting a 40% alignment on average. Napa, given the low number of samples, show zero levels of alignments, suggesting the need for more samples.



(a) kPC1-CV1.

(b) kPC2-CV1.

Figure 22: Structured coefficients of financial/climate kPC1 (left) and kPC2 (right) for the first canonical variate. These are presented in Table 22, in the fourth column.

For the MSEs given in Fig. 24, it appears that kPC2 performs better than kPC1, except for San Joaquin, where higher levels of MSEs for the second kPC can be identified. This can be because the first kPC captures more noise than the second one. The county where the biggest MSEs can be found for kPC1 is Alameda, which, however, has the greatest number of issued bonds. Napa shows deficient levels, suggesting that its kPCs mainly capture the numerical variable given the alignments described above for the categorical ones. In practice, using kPC2 leads to way lower levels of MSEs for most numerical financial variables. Given these results, comparing an MSE with a cKTA is challenging. However, lower levels of MSE imply better performances and a higher level of cKTAs indicates better performances. Given the low levels of the cKTAs, the obtained kPCs are driven more by the numerical variables than the categorical ones, suggesting a higher level of importance when the cross-correlation is analysed with the CCA.

5 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 carbon reduction of these bonds presents a considerable challenge for investors due to the lack of standardised reporting on environmental impact. Our research initiative designed a unique set of indicators to 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.

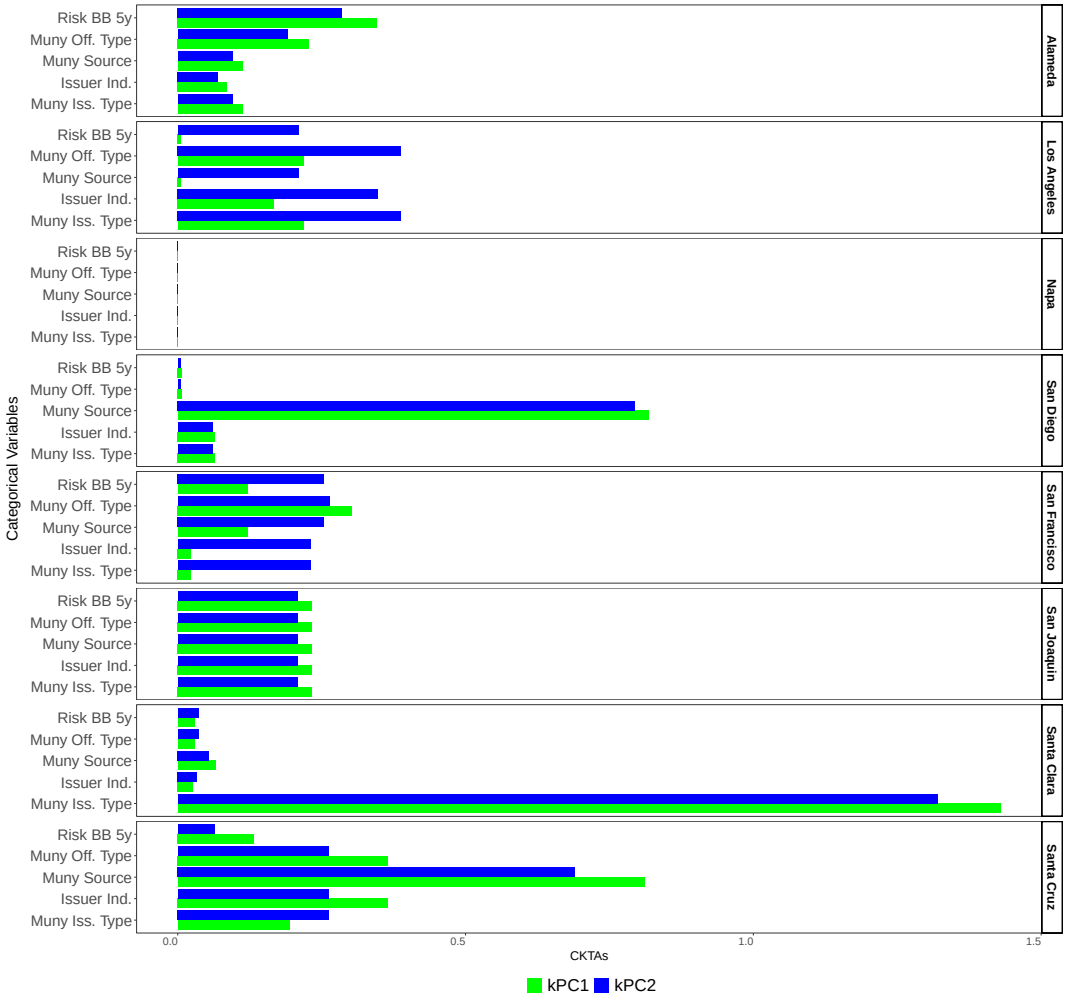


Figure 23: cKTAs of kPC1 and kPC2 by county in capturing each individual categorical variable. The alignments are computed between the empirical covariance matrix of a vector kPC and the empirical covariance matrix of a vector categorical variable. In the y-axis the categorical variables are given and the x-axis represents the cKTA. The cKTA is comprised in a range between -1 and 1.

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 multi-dimensional 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 multiple multivariate data, evaluating the impact of municipal green bonds as a whole within California by considering pollution and climate as attributes of the sought impact. Central to our research was applying kPCA and CCA to identify cross-correlation within multiple multivariate data sets. This novel approach allows for the combination of multiple 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.

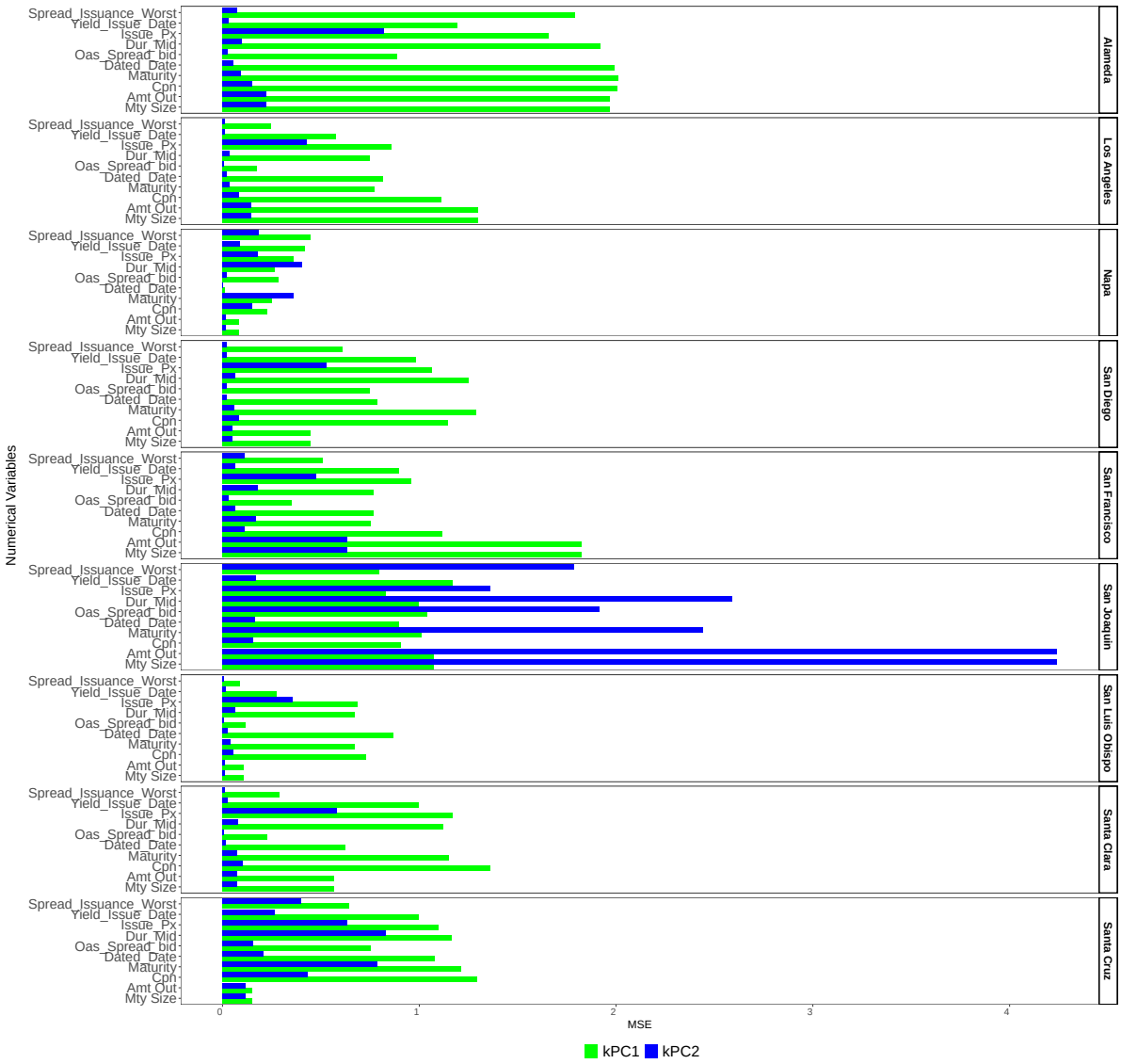


Figure 24: MSEs of kPC1 and kPC2 by county in capturing each individual numerical variable. The MSE is computed through the reconstructed pre-images obtained kPC1 and kPC1 and kPC2 and then the euclidean distances between the reconstrcuced data and the original data are computed. In the y-axis the numerical variables are given and the x-axis represents the MSEs..

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 project 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.

References

[1] Susan Solomon. *Climate change 2007-the physical science basis: Working group I contribution to the fourth assessment report of the IPCC*, volume 4. Cambridge university press, 2007.

[2] Terry L Root, Jeff T Price, Kimberly R Hall, Stephen H Schneider, Cynthia Rosenzweig, and J Alan Pounds. Fingerprints of global warming on wild animals and plants. *Nature*, 421(6918):57, 2003.

[3] Peter M Cox, Richard A Betts, Chris D Jones, Steven A Spall, and Ian J Totterdell. Acceleration of global warming due to carbon-cycle feedbacks in a coupled climate model. *Nature*, 408(6809):184, 2000.

[4] Kyoto Protocol. Kyoto. *Japan: UNFCCC, COP3*, 1997.

[5] Sebastian Oberthür and Hermann E Ott. *The Kyoto Protocol: international climate policy for the 21st century*. Springer Science & Business Media, 1999.

[6] Kyoto Protocol. United nations framework convention on climate change. *Kyoto Protocol, Kyoto*, 19, 1997.

[7] Nicholas Stern et al. What is the economics of climate change? *WORLD ECONOMICS-HENLEY ON THAMES-*, 7(2):1, 2006.

[8] Merrian C Fuller, Stephen Compagni Portis, and Daniel M Kammen. Toward a low-carbon economy: municipal financing for energy efficiency and solar power. *Environment: Science and Policy for Sustainable Development*, 51(1):22–33, 2009.

[9] Emanuele Campiglio. Beyond carbon pricing: The role of banking and monetary policy in financing the transition to a low-carbon economy. *Ecological Economics*, 121:220–230, 2016.

[10] Haruna Gujba, Steve Thorne, Yacob Mulugetta, Kavita Rai, and Youba Sokona. Financing low carbon energy access in africa. *Energy Policy*, 47:71–78, 2012.

[11] Jian-qiang BAO, Yang Miao, and Feng CHEN. Low carbon economy: Revolution in the way of human economic development [j]. *China Industrial Economics*, 4(2008):017, 2008.

[12] Koji Shimada, Yoshitaka Tanaka, Kei Gomi, and Yuzuru 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):4688–4703, 2007.

[13] Ann P Kinzig and Daniel M Kammen. National trajectories of carbon emissions: analysis of proposals to foster the transition to low-carbon economies. *Global Environmental Change*, 8(3):183–208, 1998.

[14] Nicholas Stern. Stern review report on the economics of climate change. 2006.

[15] Stephany Griffith-Jones and Judith Tyson. *The European Investment Bank: Lessons for Developing Countries*. Number 2013/019. WIDER Working Paper, 2013.

[16] International Capital Market Association et al. Green bond principles: voluntary process guidelines for issuing green bonds. *International Capital Market Association, Zürich*, 2018.

[17] International Capital Market Association et al. Green bond principles. *Retrieved from International Capital Market Association website: http://www. icmagroup. org/Regulatory-Policy-and-Market-Practice/green-bonds/green-bond-principles*, 2014.

[18] American Lung Association. State of the air. *American Lung Association*, 2023.

[19] Leah Fisher and Sonya Ziaja. Statewide summary report. *California's Fourth Climate Assessment*, 2018.

[20] Bernhard Schölkopf, Alexander Smola, and Klaus-Robert Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural computation*, 10(5):1299–1319, 1998.

[21] Wolfgang Karl Härdle and Léopold Simar. Canonical correlation analysis. In *Applied multivariate statistical analysis*, pages 443–454. Springer, 2015.

[22] Harold Hotelling. The most predictable criterion. *Journal of educational Psychology*, 26(2):139, 1935.

[23] Ian T Joliffe and BJT Morgan. Principal component analysis and exploratory factor analysis. *Statistical methods in medical research*, 1(1):69–95, 1992.

[24] Vladimir Vapnik. *The nature of statistical learning theory*. Springer science & business media, 1999.

[25] Francis R Bach and Michael I Jordan. Kernel independent component analysis. *Journal of machine learning research*, 3(Jul):1–48, 2002.

[26] David R Hardoon, Sandor Szedmak, and John Shawe-Taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural computation*, 16(12):2639–2664, 2004.

[27] Kosuke Yoshida, Junichiro Yoshimoto, and Kenji Doya. Sparse kernel canonical correlation analysis for discovery of nonlinear interactions in high-dimensional data. *BMC bioinformatics*, 18:1–11, 2017.

[28] Weiran Wang and Karen Livescu. Large-scale approximate kernel canonical correlation analysis. *arXiv preprint arXiv:1511.04773*, 2015.

[29] David Lopez-Paz, Suvrit Sra, Alex Smola, Zoubin Ghahramani, and Bernhard Schölkopf. Randomized nonlinear component analysis. In *International conference on machine learning*, pages 1359–1367. PMLR, 2014.

[30] Viivi Uurtio, Sahely Bhadra, and Juho Rousu. Sparse non-linear cca through hilbert-schmidt independence criterion. In *2018 IEEE International Conference on Data Mining (ICDM)*, pages 1278–1283. IEEE, 2018.

[31] Viivi Uurtio, Sahely Bhadra, and Juho Rousu. Large-scale sparse kernel canonical correlation analysis. In *International Conference on Machine Learning*, pages 6383–6391. PMLR, 2019.

[32] Jinglin Xu, Wenbin Li, Xinwang Liu, Dingwen Zhang, Ji Liu, and Junwei Han. Deep embedded complementary and interactive information for multi-view classification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 6494–6501, 2020.

[33] Natalia Y Bilenko and Jack L Gallant. Pyrcca: regularized kernel canonical correlation analysis in python and its applications to neuroimaging. *Frontiers in neuroinformatics*, 10:49, 2016.

[34] Paul Honeine and Cédric Richard. A closed-form solution for the pre-image problem in kernel-based machines. *Journal of Signal Processing Systems*, 65(3):289–299, 2011.

[35] Gökhan H Bakır, Jason Weston, and Bernhard Schölkopf. Learning to find pre-images. *Advances in neural information processing systems*, 16:449–456, 2004.

[36] Craig Saunders, Alexander Gammerman, and Volodya Vovk. Ridge regression learning algorithm in dual variables. 1998.

[37] Patrick V Dattalo. A demonstration of canonical correlation analysis with orthogonal rotation to facilitate interpretation. 2014.

[38] Alissa Sherry and Robin K Henson. Conducting and interpreting canonical correlation analysis in personality research: A user-friendly primer. *Journal of personality assessment*, 84(1):37–48, 2005.

[39] Mark S Levine. *Canonical analysis and factor comparison*. Number 6. Sage, 1977.

[40] Calyampudi Radhakrishna Rao, Calyampudi Radhakrishna Rao, Mathematischer Statistiker, Calyampudi Radhakrishna Rao, and Calyampudi Radhakrishna Rao. *Linear statistical inference and its applications*, volume 2. Wiley New York, 1973.

[41] Troy Courville and Bruce Thompson. Use of structure coefficients in published multiple regression articles: β is not enough. *Educational and Psychological Measurement*, 61(2):229–248, 2001.

[42] Robin K Henson. The logic and interpretation of structure coefficients in multivariate general linear model analyses. 2002.

[43] Andreas Alfons, Christophe Croux, and Peter Filzmoser. Robust maximum association estimators. *Journal of the American Statistical Association*, 112(517):436–445, 2017.

[44] Bart Rousseau, Stefan T Maes, and Albert TH Lenstra. Systematic intensity errors and model imperfection as the consequence of spectral truncation. *Acta Crystallographica Section A: Foundations of Crystallography*, 56(3):300–307, 2000.

[45] Paul Jaccard. Distribution de la flore alpine dans le bassin des dranses et dans quelques régions voisines. *Bull Soc Vaudoise Sci Nat*, 37:241–272, 1901.

[46] Luciano da F Costa. Further generalizations of the jaccard index. *arXiv preprint arXiv:2110.09619*, 2021.

[47] Peter Molnár. High-low range in garch models of stock return volatility. *Applied Economics*, 48(51):4977–4991, 2016.

[48] Patricio Cerda, Gaël Varoquaux, and Balázs Kégl. Similarity encoding for learning with dirty categorical variables. *Machine Learning*, 107(8-10):1477–1494, 2018.

[49] Corinna Cortes, Mehryar Mohri, and Afshin Rostamizadeh. Two-stage learning kernel algorithms. 2010.

[50] Christopher JC Burges et al. Simplified support vector decision rules. In *ICML*, volume 96, pages 71–77. Citeseer, 1996.

[51] Sebastian Mika, Bernhard Schölkopf, Alex J Smola, Klaus-Robert Müller, Matthias Scholz, and Gunnar Rätsch. Kernel pca and de-noising in feature spaces. In *Advances in neural information processing systems*, pages 536–542, 1999.

[52] Md Ashad Alam and Kenji Fukumizu. Hyperparameter selection in kernel principal component analysis. 2014.

[53] JT-Y Kwok and IW-H Tsang. The pre-image problem in kernel methods. *IEEE transactions on neural networks*, 15(6):1517–1525, 2004.

[54] Trine Julie Abrahamsen and Lars Kai Hansen. Input space regularization stabilizes pre-images for kernel pca de-noising. In *2009 IEEE International Workshop on Machine Learning for Signal Processing*, pages 1–6. IEEE, 2009.

[55] Gökhan Bakir. Extensions to kernel dependency estimation-with applications to robotics. 2006.

[56] Paul Honeine and Cedric Richard. Preimage problem in kernel-based machine learning. *IEEE Signal Processing Magazine*, 28(2):77–88, 2011.

[57] Wei-Shi Zheng and Jian-huang Lai. Regularized locality preserving learning of pre-image problem in kernel principal component analysis. In *18th International Conference on Pattern Recognition (ICPR'06)*, volume 2, pages 456–459. IEEE, 2006.

[58] Wei-Shi Zheng, JianHuang Lai, and Pong C Yuen. Penalized preimage learning in kernel principal component analysis. *IEEE Transactions on Neural Networks*, 21(4):551–570, 2010.

Authors' Contributions

Dr. Marta Campi: Co-developed methodology, developed all implementation and results, obtained and prepared data, wrote all code, co-wrote first draft.

Chair Prof. Gareth W. Peters: Developed problem methodology, original mathematical problem derivation and draft, co-developed applications plan and co-wrote first draft.

Dr. Kylie-Anne Richards: Co-developed the applications, obtained financial data and developed financial context and details, co-wrote first draft.

A Alternatives Methods for the Pre-Image Problem

A.0.1 Solving The Pre-Image Problem

A problem is ill posed if at least one of the following three conditions, which characterised well-posed problems in the sense of Hadamard, is violated: (i) a solution exists, (ii) it is unique, and (iii) it depends continuously on the data (also known as the stability condition). Unfortunately, identifying the pre-image is generally an ill-posed problem. This is the result of $\dim(\mathcal{H}) \gg \dim(\mathcal{X})$ and the fact that φ is not surjective. Remark that $\varphi : \mathcal{X} \to \mathcal{H}$ is a possibly non-linear map from the D-dimensional input space $\mathcal{X}$ to the high (possibly infinite) dimensional feature space $\mathcal{H}$. Furthermore, whether φ is injective depends on the choice of kernel function. As a function $f : X \to Y$ has an inverse iff it is bijective, we do not expect φ to have an inverse. When φ is not surjective, it follows that not all points in $\mathcal{H}$ or even the span of $\{\varphi(\mathcal{X})\}$ is the image of some $\mathbf{x}^\star \in \mathcal{X}$. Finally, when φ is not injective any recovered image might not be unique. This means that there may not exist $\mathbf{x}^\star$ such that $\varphi(\mathbf{x}^\star) = \phi$. In order to circumvent this difficulty, one seeks an approximate solution, i.e. $\mathbf{x}^\star$ whose map $\varphi(\mathbf{x}^\star)$ is as close as possible to ϕ.
Consider a pattern ϕ in the feature space $\mathcal{H}$, obtained by any kernel-based machine, e.g. a principal axe or a denoised pattern from kernel-PCA. By virtue of the representer theorem, let $\phi = \sum_{i=1}^{N} a_i \varphi(\mathbf{x}_i)$. The pre-image problem consists of the following optimization problem:

$$\mathbf{x}^\star = \text{argmin}_{\mathbf{x}\in\mathcal{X}} \left\| \sum_{i=1}^{N} a_i \varphi(\mathbf{x}_i) - \varphi(\mathbf{x}) \right\| \tag{45}$$

Equivalently, from the kernel trick, $\mathbf{x}^\star$, minimizes the objective function

$$\rho(\mathbf{x}) = k(\mathbf{x}, \mathbf{x}) - 2 \sum_{i=1}^{N} a_i k(\mathbf{x}, \mathbf{x}_i) \tag{46}$$

where the term independente of $\mathbf{x}$ has been dropped. As opposed to this functional formalism, one may also adopt a vector-wise representation, with elements in the RKHS given by their coordinates with respect to an orthogonal basis. Taking for instance the basis defined by the kernel-PCA, as given in **??**, each $\phi \in \mathcal{H}$ is represented vector-wise with $[\langle \phi, \phi_1 \rangle, \langle \phi, \phi_2 \rangle, \ldots, \langle \phi, \phi_\ell \rangle]^\intercal$, thus defining an ℓ-dimensional representation. In such a case, the Euclidean distance between the latter and the one obtained from the image of $\mathbf{x}^\star$ is minimized. This is a classical reduction problem, connecting the pre-image problem to the historical evolution of dimensionality reduction techniques. The problem in 45 is inherently non-linear and non-convex for many choices of kernel function, making it non-trivial to find a reliable pre-image. The solutions used so far to solve this non-linear optimization problem often employ gradient descent or non-linear iteration methods (see amongst others [50], [51]). We review some of these methods in the subsections below.

A.0.2 The Exact Pre-Image, When It Exists

Suppose for now that there exist an exact pre-image of ϕ, i.e. $\mathbf{x}^\star$ such that $\varphi(\mathbf{x}^\star) = \phi$, then the optimisation problem in 45 results into that pre-image. Furthermore, the pre-image can be easily computed when the kernel is an invertible function of $\langle \mathbf{x}_i, \mathbf{x}_j \rangle$, such as some projective kernels including the polynomial kernel with odd degree and the sigmoid kernel. Let $h : \mathbb{R} \to \mathbb{R}$ defines the inverse function such that $h(k(\mathbf{x}_i, \mathbf{x}_j)) = \langle \mathbf{x}_i, \mathbf{x}_j \rangle$. Then, given any orthonormal basis in the input space $\{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_N\}$, every element $\mathbf{x} \in \mathcal{X}$ can be written as

$$\mathbf{x} = \sum_{j=1}^{N} \langle \mathbf{e}_j, \mathbf{x} \rangle \mathbf{e}_j = \sum_{j=1}^{N} h(k(\mathbf{e}_j, \mathbf{x})) \mathbf{e}_j \tag{47}$$

As a consequence, the exact pre-image $\mathbf{x}^\star$ of some pattern $\phi = \sum_{i=1}^N a_i \varphi(\mathbf{x}_i)$, namely $\varphi(\mathbf{x}^\star) = \phi$, can be expanded as

$$\mathbf{x}^\star = \sum_{j=1}^N h\left(\sum_{i=1}^N a_i k(\mathbf{e}_j, \mathbf{x}_i)\right)\mathbf{e}_j \tag{48}$$

Likewise, when the kernel is an invertible function of the distance, such as radial kernels, a similar expression can be derived by using the polarization identity $4\langle \mathbf{x}^\star, \mathbf{e}_j\rangle = \|\mathbf{x}^\star + \mathbf{e}_j\| - \|\mathbf{x}^\star - \mathbf{e}_j\|$. Clearly, such a simple derivation for the pre-image is only valid under the crucial assumption that the pre-image $\mathbf{x}^\star$ exists. Unfortunately, for a large class of kernels, there are no exact pre-images. Rather than seeking the exact pre-image, we consider an approximate pre-image by solving the optimisation problem in 45. We review some of the most common employed techniques to solve such an optimisation problem within the next paragraphs.

A.0.3 Gradient Descent Techniques

Gradient descent is one of the simplest optimization techniques. It requires computing the gradient of the objective function 46, which we denote $\nabla_{\mathbf{x}}\rho(\mathbf{x}^\star)$. In its simplest form, the current guess $\mathbf{x}_t^\star$ is updated into $\mathbf{x}_{t+1}^\star$ by stepping into the direction opposite to the gradient, with

$$\mathbf{x}_{t+1}^\star = \mathbf{x}_t^\star - \eta_t \nabla_{\mathbf{x}} J(\mathbf{x}_t^\star) \tag{49}$$

where η_t is a step size parameter, often optimized by using a line-search procedure. As an alternative to the gradient descent, one may use more sophisticated techniques, such as Newton's method. Unfortunately, the objective function is non-linear and non-convex. Thus, a gradient descent algorithm must be run many times with several different starting values, in hope that a feasible solution will be amongst the local minima obtained over the runs.

A.0.4 Fixed-Point Iteration Method

The structure of kernel functions provides useful insights to derive more appropriate optimization techniques, beyond classical gradient descent. More precisely, the gradient of expression 46 has a closed-form expression for most kernels. By setting this expression to zero, this greatly simplifies the optimization scheme, resulting into a fixed-point iterative technique. Taking for instance the Gaussian kernel [52] and [51], the objective function in 46 becomes

$$-2\sum_{i=1}^N a_i \exp\left(-\|\mathbf{x} - \mathbf{x}_i\|^2 / 2\sigma^2\right) \tag{50}$$

with its gradient

$$\nabla_{\mathbf{x}}\rho(\mathbf{x}) = -\frac{2}{\sigma^2}\sum_{i=1}^N a_i \exp\left(-\|\mathbf{x} - \mathbf{x}_i\|^2 / 2\sigma^2\right)(\mathbf{x} - \mathbf{x}_i) \tag{51}$$

We get the pre-image by setting this gradient to zero, which results into the fixed-point iterative expression

$$\mathbf{x}_t^\star = \frac{\sum_{i=1}^N a_i k(\mathbf{x}_t^\star, \mathbf{x}_i)\mathbf{x}_i}{\sum_{i=1}^N a_i k(\mathbf{x}_t^\star, \mathbf{x}_i)} \tag{52}$$

with $k(\mathbf{x}_t^\star, \mathbf{x}_i) = \exp\left(-\|\mathbf{x} - \mathbf{x}_i\|^2 / 2\sigma^2\right)$. Similar expressions can be derived for most kernels, such as the polynomial kernel of degree p (see [53]) with

$$\mathbf{x}_t^\star = \sum_{i=1}^N a_i \left(\frac{\langle \mathbf{x}_t^\star, \mathbf{x}_i\rangle + c}{\langle \mathbf{x}_t^\star, \mathbf{x}_t^\star\rangle + c}\right)^{p-1}\mathbf{x}_i \tag{53}$$

Unfortunately, the fixed-point iterative technique still suffers from local minima and tends to be unstable. The numerical instability occurs especially when the value of the denominator decreases to zero. To prevent this situation, a regularized solution can be easily formulated, as studied in [54].

A.1 Learning the pre-image map

To find the pre-image map, a learning machine is constructed with training elements from the feature space and estimated values in the input space as follows: we seek to estimate a function $\Gamma^\star$ approximating φ^{-1} with the property that $\Gamma^\star(\varphi(\mathbf{x}_i)) = \mathbf{x}_i$ for $i = 1, \ldots, N$. Then, ideally, $\Gamma^\star(\phi)$ should give $\mathbf{x}^\star$, the pre-image of ϕ. To make the computationally tractable, two issues are considered in [35], [55] and reviewed in [56]. First, the function is defined on a vector space. This can be done by representing a vector-wise any $\phi \in \mathcal{H}$ with $[\langle \phi, \phi_1 \rangle, \ldots, \langle \phi, \phi_\ell \rangle]^\intercal$, using an orthogonal basis obtained from kernel-PCA. Second, the pre-image map $\Gamma^\star$ is decomposed into $\dim(\mathcal{X})$ functions to estimate each component of $\mathbf{x}^\star$. From these considerations, we seek functions $\Gamma_1^\star, \Gamma_2^\star, \ldots, \Gamma_{\dim(\mathcal{X})}^\star$ with $\Gamma_j^\star : \mathbb{R}^\ell \to \mathbb{R}$. Each of these functions is obtained by solving the optimisation problem

$$\Gamma_j^\star = \text{argmin}_\Gamma \sum_{i=1}^{N} f([\mathbf{x}_i]_j, \Gamma(\phi)) + \eta g(\|\Gamma\|^2) \tag{54}$$

where $f(\cdot, \cdot)$ is some loss function, and $[\cdot]_j$ denotes the j-th component operator. By taking for instance the distance as a loss function, we obtain

$$\Gamma_j^\star = \text{argmin}_\Gamma \frac{1}{N} \sum_{i=1}^{N} \left\| [\mathbf{x}_i]_j - \Gamma(\phi) \right\|^2 + \eta \|\Gamma\|^2 \tag{55}$$

This optimization problem can be easily solved by a matrix inversion scheme, in analogy to the ridge regression problem **??** and its linear system **??**. Several enhancements of this method have been proposed, see [57] and [58]. All these methods are based on a set of available data in the input space and the associated images in the RKHS. What is important to highlight at this stage is the observation that a convenient way of working in the subspace (of finite dimension) defined by the data set spanning $\mathcal{H}$ is indeed considering a set of coordinates, e.g. the kernel-PCA basis. Consider $P_\ell \varphi(\mathbf{x})$, where $P_\ell = \left(\sum_{i=1}^N a_i \varphi(\mathbf{x}_1), \ldots, \sum_{i=1}^N a_i \varphi(\mathbf{x}_\ell) \right)$, defined as

$$P_\ell \varphi(\mathbf{x}) = \sum_{k=1}^{\ell} \mathbf{y}_k(\mathbf{x}) \phi_k = \sum_{k=1}^{\ell} \langle \phi_k, \varphi(\mathbf{x}) \rangle \phi_k = \sum_{k=1}^{\ell} \frac{1}{\sqrt{\lambda_k}} \sum_{i=1}^{N} a_{ki} k(\mathbf{x}, \mathbf{x}_i) \phi_k \tag{56}$$

It represents the projection that maps a feature vector $\varphi(\mathbf{x})$ into its coordinates in the PCA basis $\phi_1, \ldots, \phi_\ell$ i.e. into the subspace where the training set has non-zero variance. Such a projection implies the selection of a kernel k often referred to as push-forward kernel which is embedded in the expression of $P_\ell \varphi(\mathbf{x})$. The identification of the optimal kernel k with respect to the performances of the kPCA will be directly related to the choice of its hyperparameters. These will be selected according to the the best performances shown by the approximated pre-images, given that within the input space we can then compare the euclidean distances, while in the feature space, the RKHS norms will not be comparable. The model assumed for the pre-image mapping $\Gamma(\phi)$ given as $\Gamma(P_\ell \varphi(\mathbf{x}))$ is an additive model defined through a second kernel κ, which differs from k, defined as follows:

$$\Gamma(P_\ell \varphi(\mathbf{x})) = \sum_{i=1}^{N} \beta_i^j \kappa(P_\ell \varphi(\mathbf{x}), P_\ell \varphi(\mathbf{x}_i)) \tag{57}$$

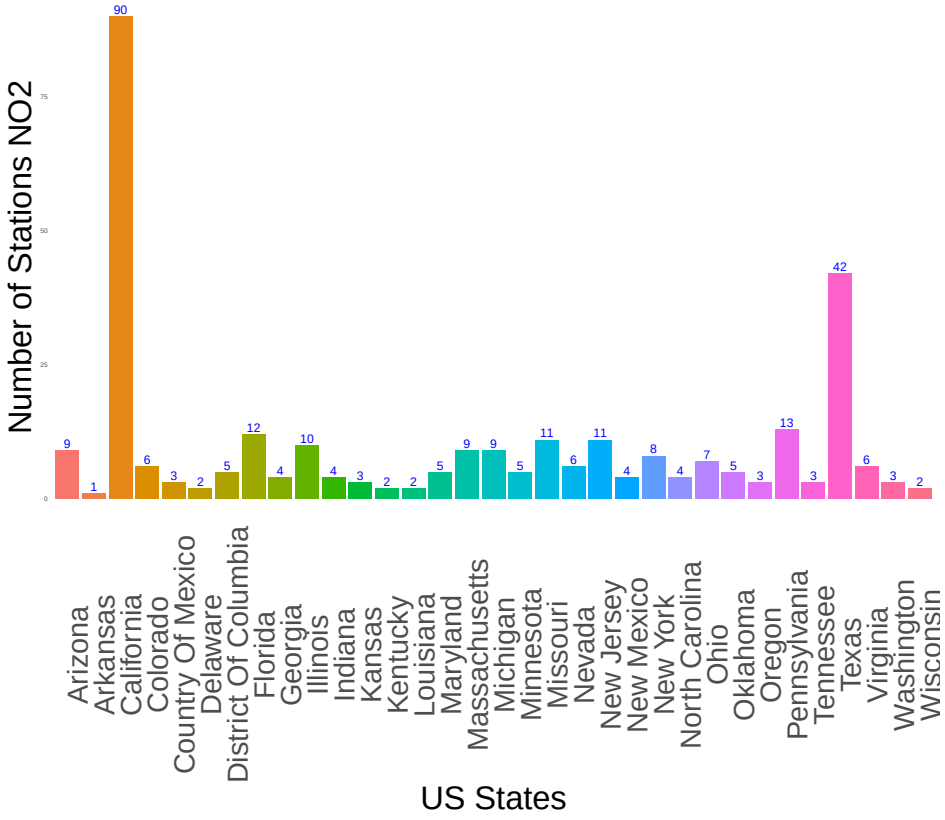
for $j = 1, \ldots, d$ or ℓ. The idea of employing a different kernel κ again is given by the fact that we want to characterise functions of the form $P_\ell \varphi(\mathbf{x})$. The main point in this step is that such a kernel will be common across all the k in terms of both kernel family and hyperparameters, which will be fixed. As a consequence, we can then compare all the obtained pre-images given that they share the same norm associated to $\kappa(\cdot, \cdot)$ and hence lying in the same space. As a result, the above optimisation problem will become

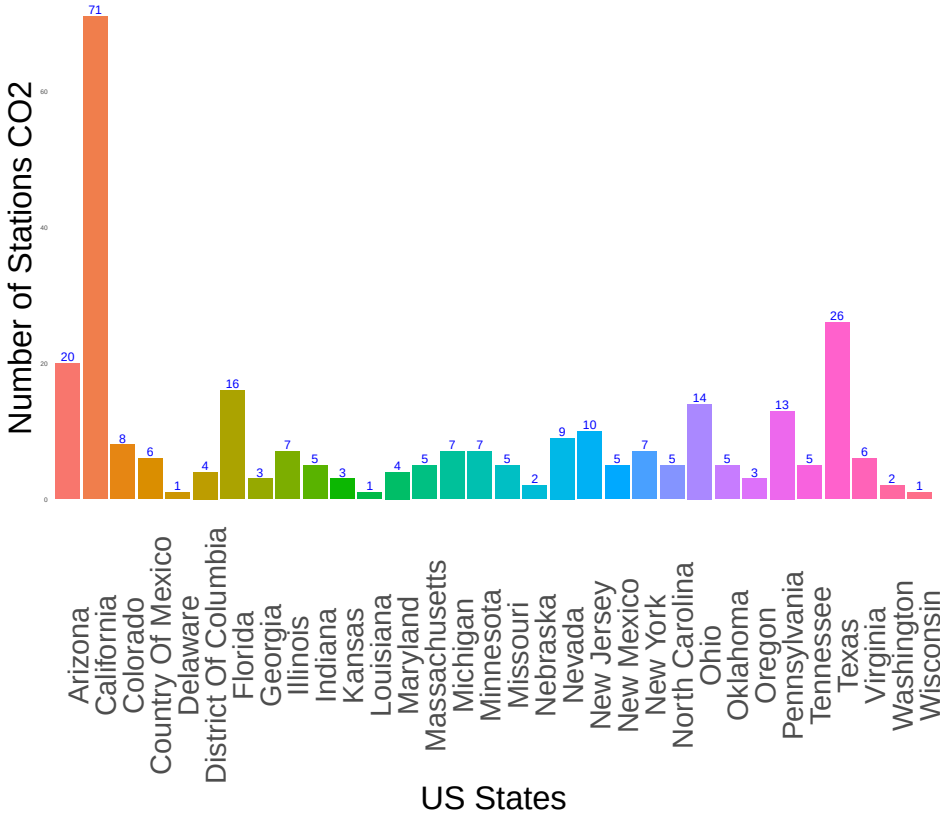
$$\beta_j^\star = \text{argmin}_{\beta^j} \frac{1}{N} \sum_{i=1}^{N} \|\varphi(\mathbf{x}_i)^\intercal \phi_j - \Gamma(P_\ell \varphi(\mathbf{x}))\|^2 + \gamma \left\| \beta^j \right\|^2 \tag{58}$$

where $\gamma \left\| \beta^j \right\|^2$ acts as regularisation term (with $\gamma > 0$), $\Gamma(P_\ell \varphi(\mathbf{x}))$ is above defined, and $\beta \in \mathbb{R}^{N \times \ell}$. Let $\mathbf{P} \in \mathbb{R}^{N \times \ell}$ with $P_{ij} = \varphi(\mathbf{x}_i)^\intercal \phi_j$ with $j = 1, \ldots, \ell$ and $\mathbf{K} \in \mathbb{R}^{N \times N}$ the kernel matrix with entry $K_{st} = \kappa(\mathbf{x}_s, \mathbf{x}_t)$. It can be solved for example through kernel ridge regression, yielding to

$$\beta = (\mathbf{K}^\intercal \mathbf{K} + \gamma \mathbf{I})^{-1} \mathbf{K} \mathbf{P} \tag{59}$$

B Available Stations in US for Climate and Pollution Data Sets





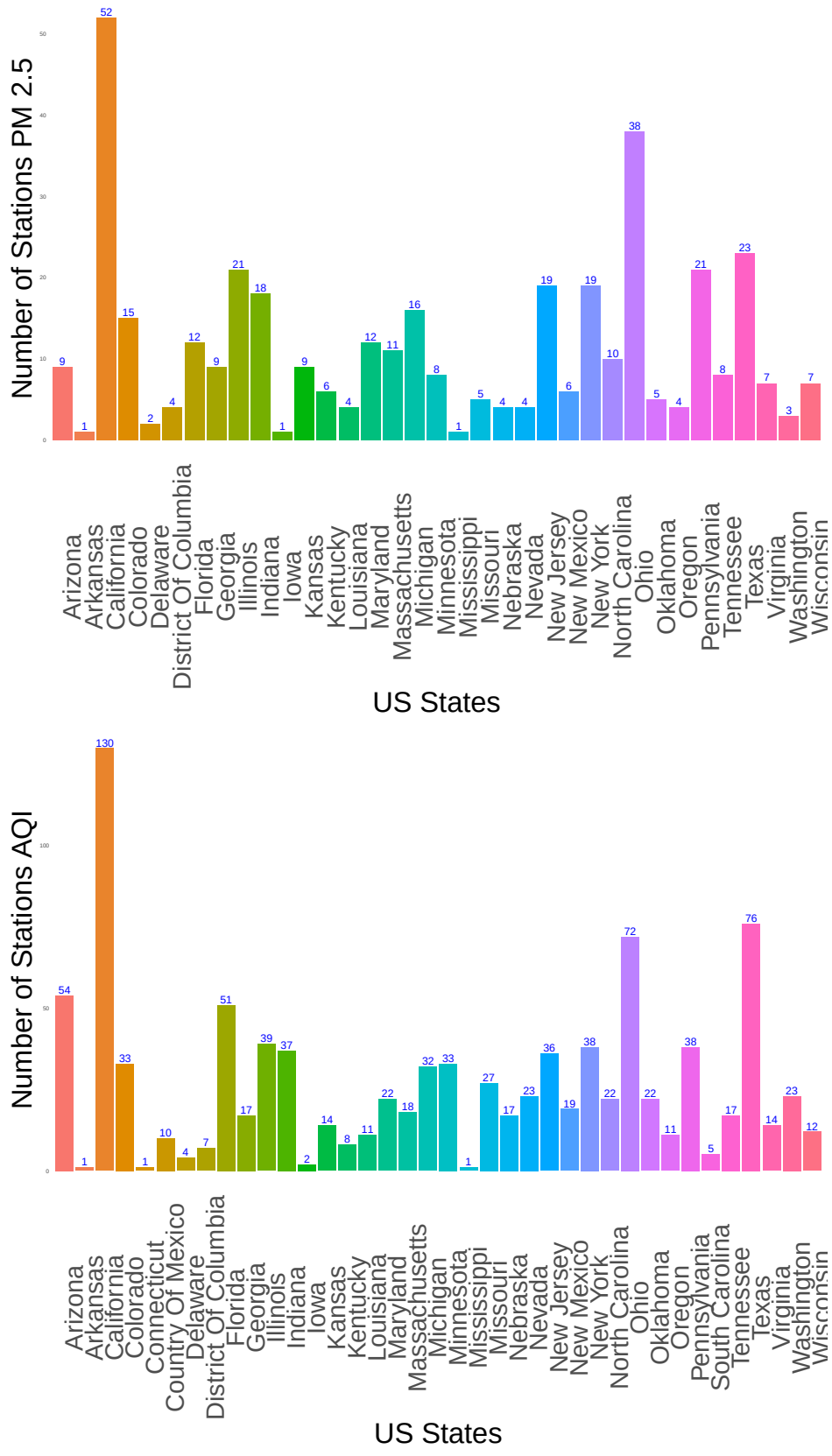


Figure 25: Available stations in the US for Pollution and Climate. This shows how the number of available stations in California are a lot bigger than in the rest of the US.

C Heatmaps by Quarters

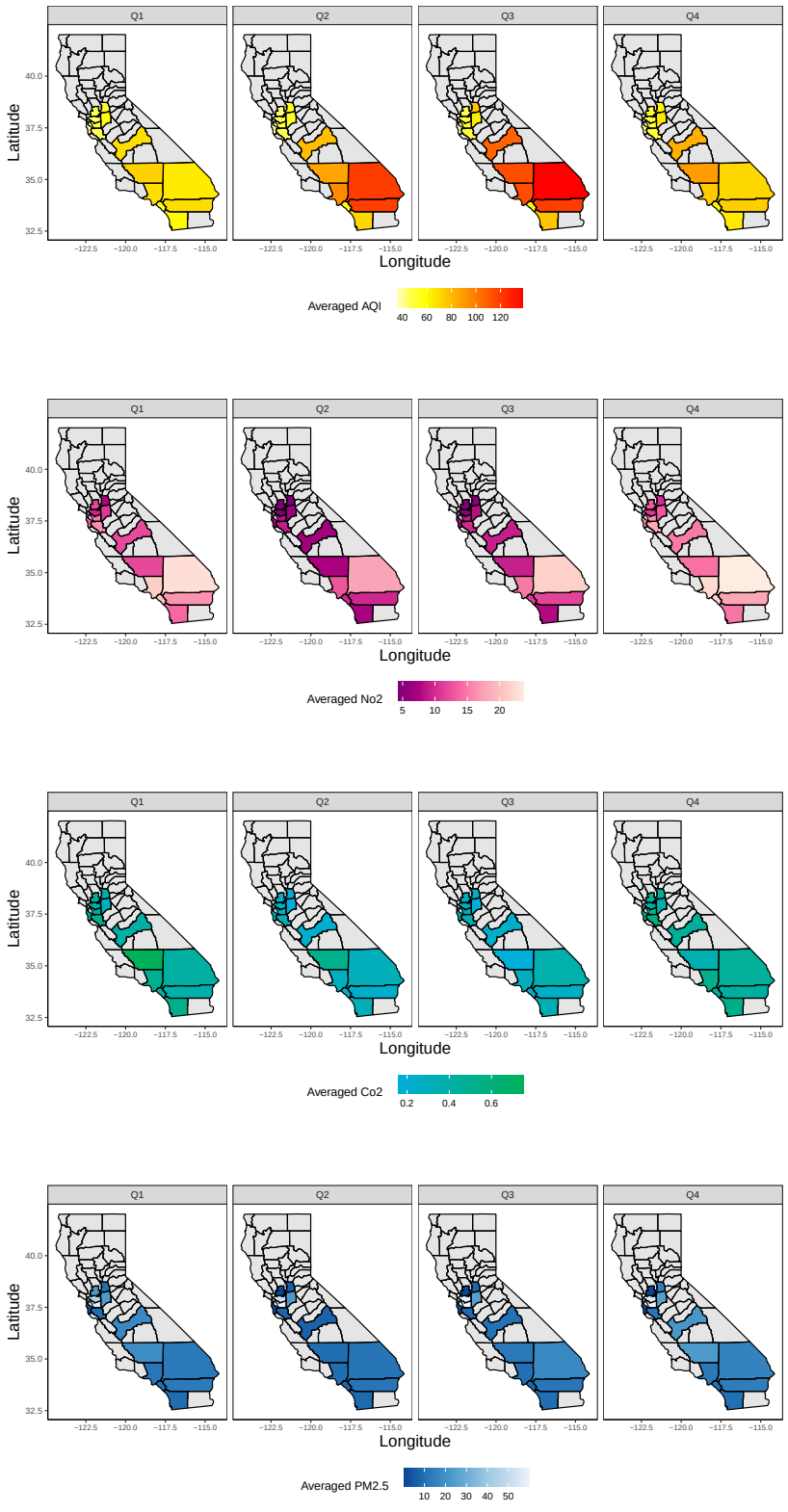


Figure 26: Heatmaps of the spatial averaged variables further averaged across each of the four quarters (across the 10 years) . In the top panels there are..

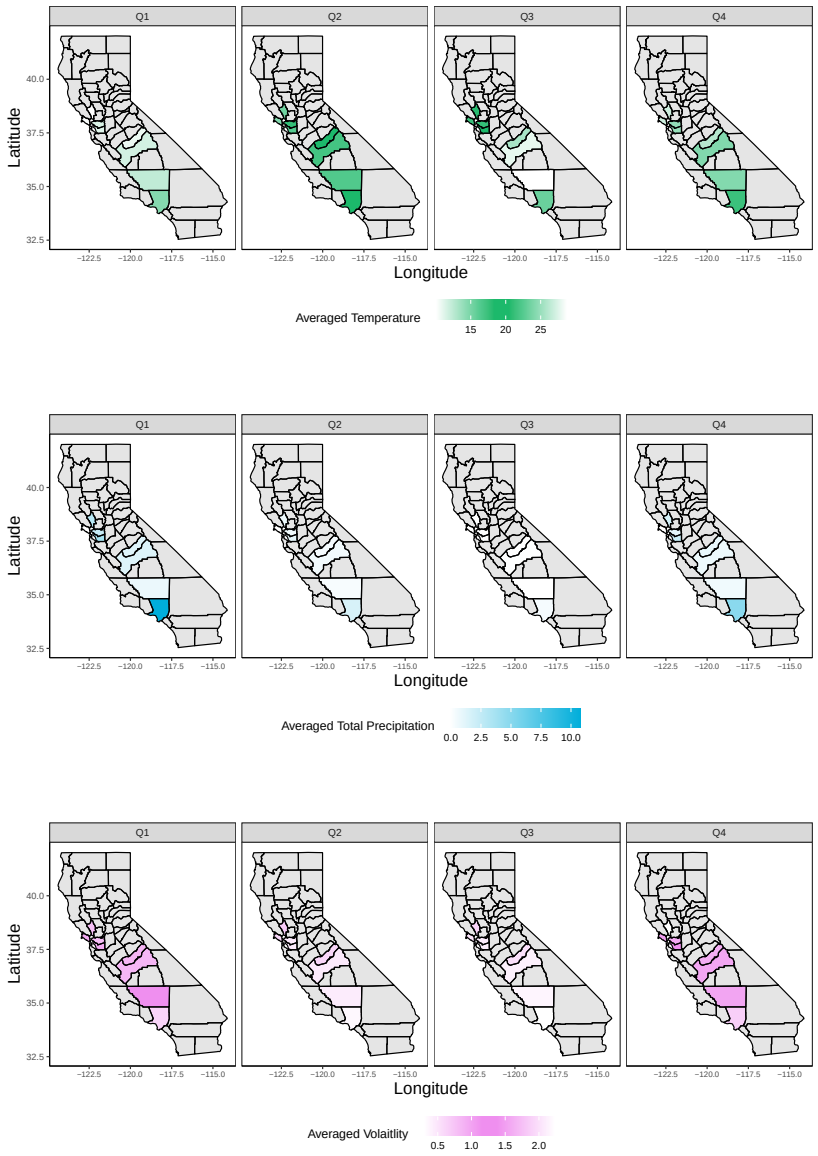


Figure 27: Heatmaps of the spatial averaged variables further averaged across each of the four quarters (across the 10 years) . In the top panels there are..

D More on the Financial Data Set

This Appendix presents further information related to the financial data set. Firstly, the number of issuers in the US, and, in particular, the great number of issuers in the State of California, represents one of the motivations for us to focus on this State for the case study of this paper. Afterwards, we present further plots describing the number of issued bonds split by county and how these are split according to some variables.

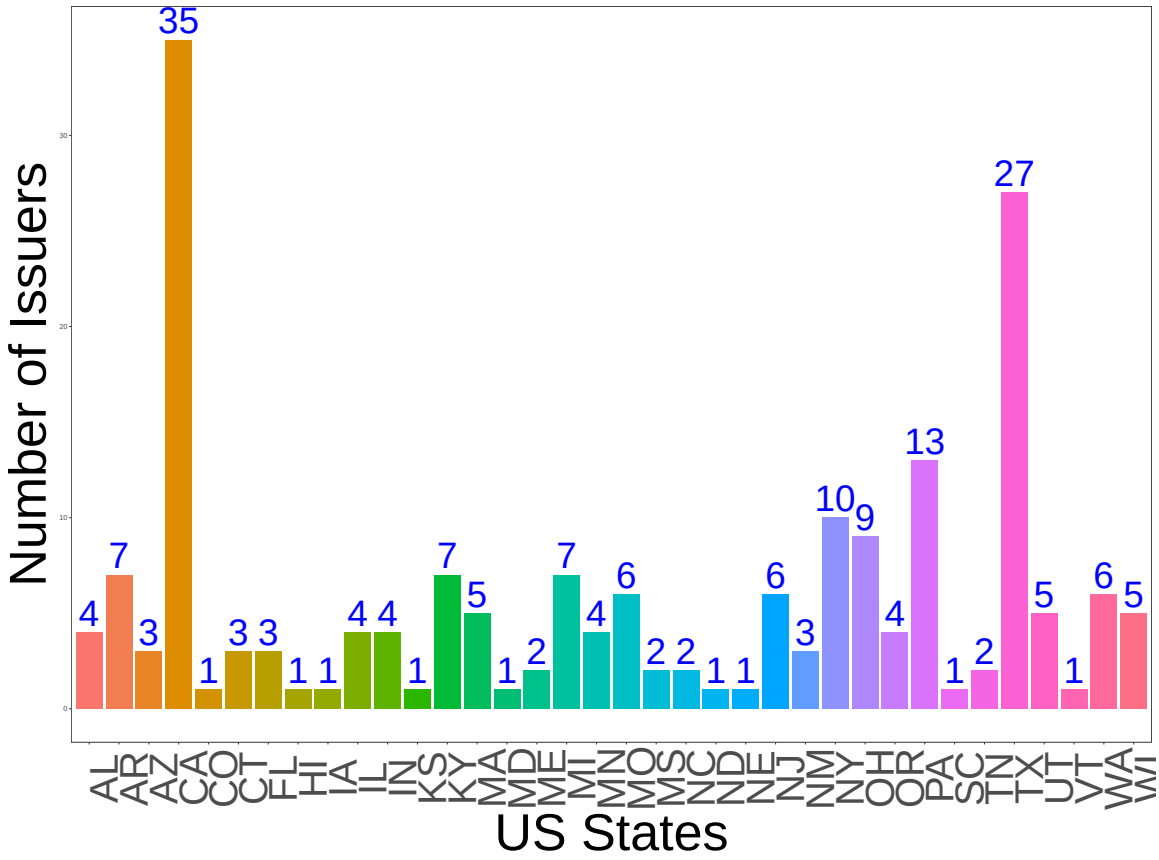
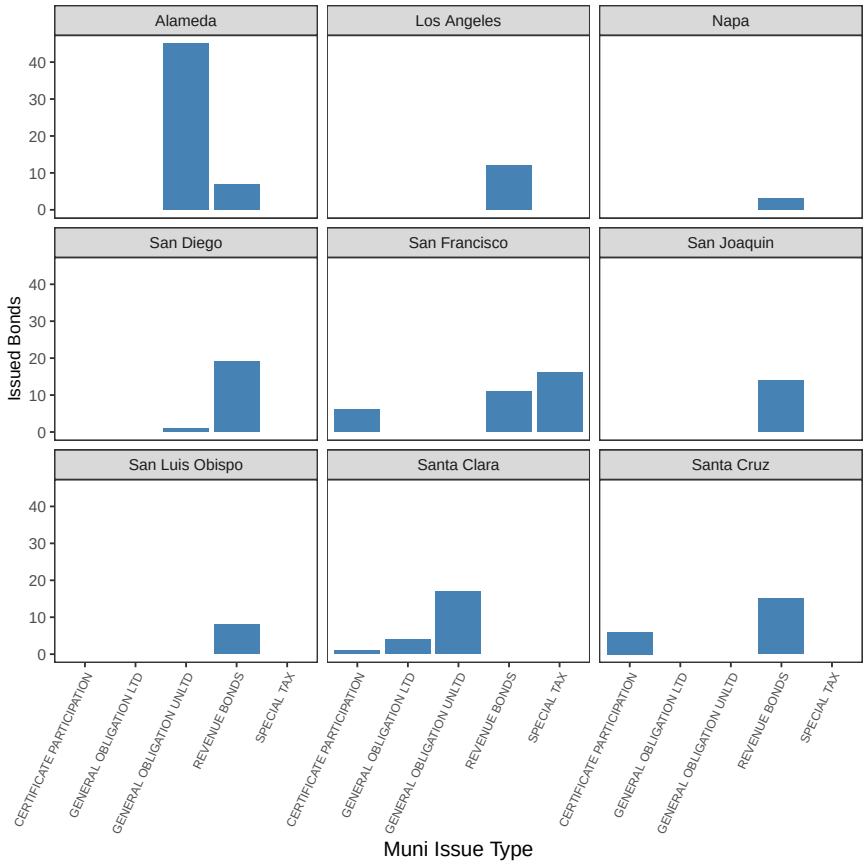
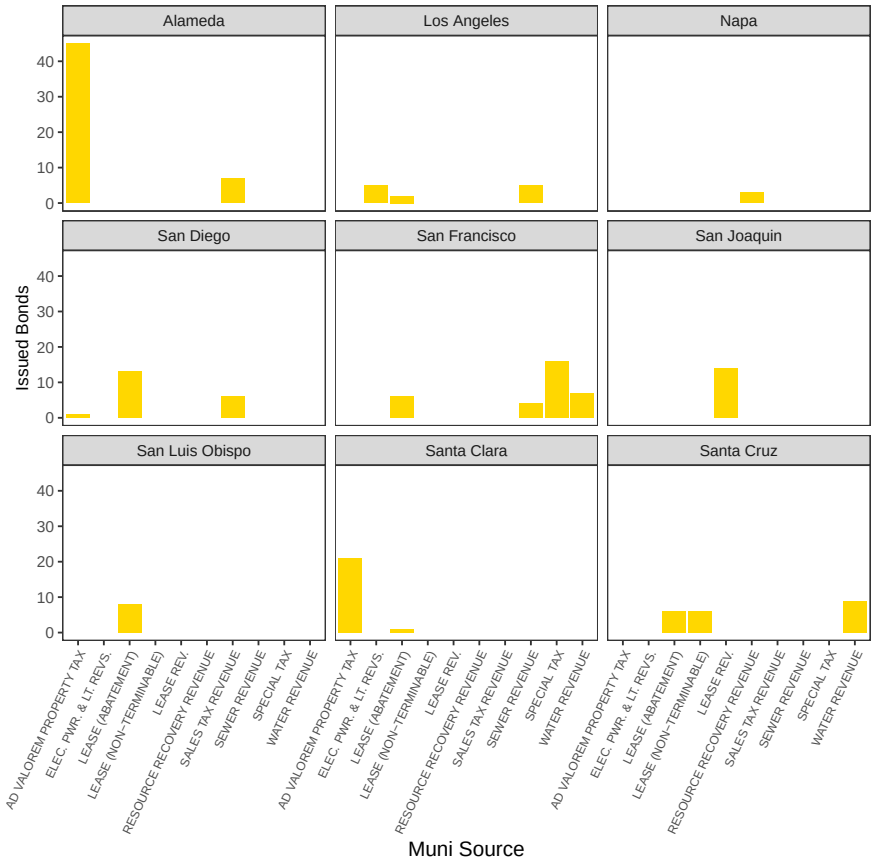


Figure 28: Number of issuers in US.


Alameda
Los Angeles
Napa
San Diego
San Francisco
San Joaquin
San Luis Obispo
Santa Clara
Santa Cruz
Issued Bonds
0
10
20
30
40
CERTIFICATE PARTICIPATION
GENERAL OBLIGATION LTD
GENERAL OBLIGATION UNLTD
REVENUE BONDS
SPECIAL TAX
Muni Issue Type


Alameda
Los Angeles
Napa
San Diego
San Francisco
San Joaquin
San Luis Obispo
Santa Clara
Santa Cruz
Issued Bonds
0
10
20
30
40
AD VALOREM PROPERTY TAX
ELEC. PWR. & LT. REVS.
LEASE (ABATEMENT)
LEASE (NON−TERMINABLE)
LEASE REV.
RESOURCE RECOVERY REVENUE
SALES TAX REVENUE
SEWER REVENUE
SPECIAL TAX
WATER REVENUE
Muni Source

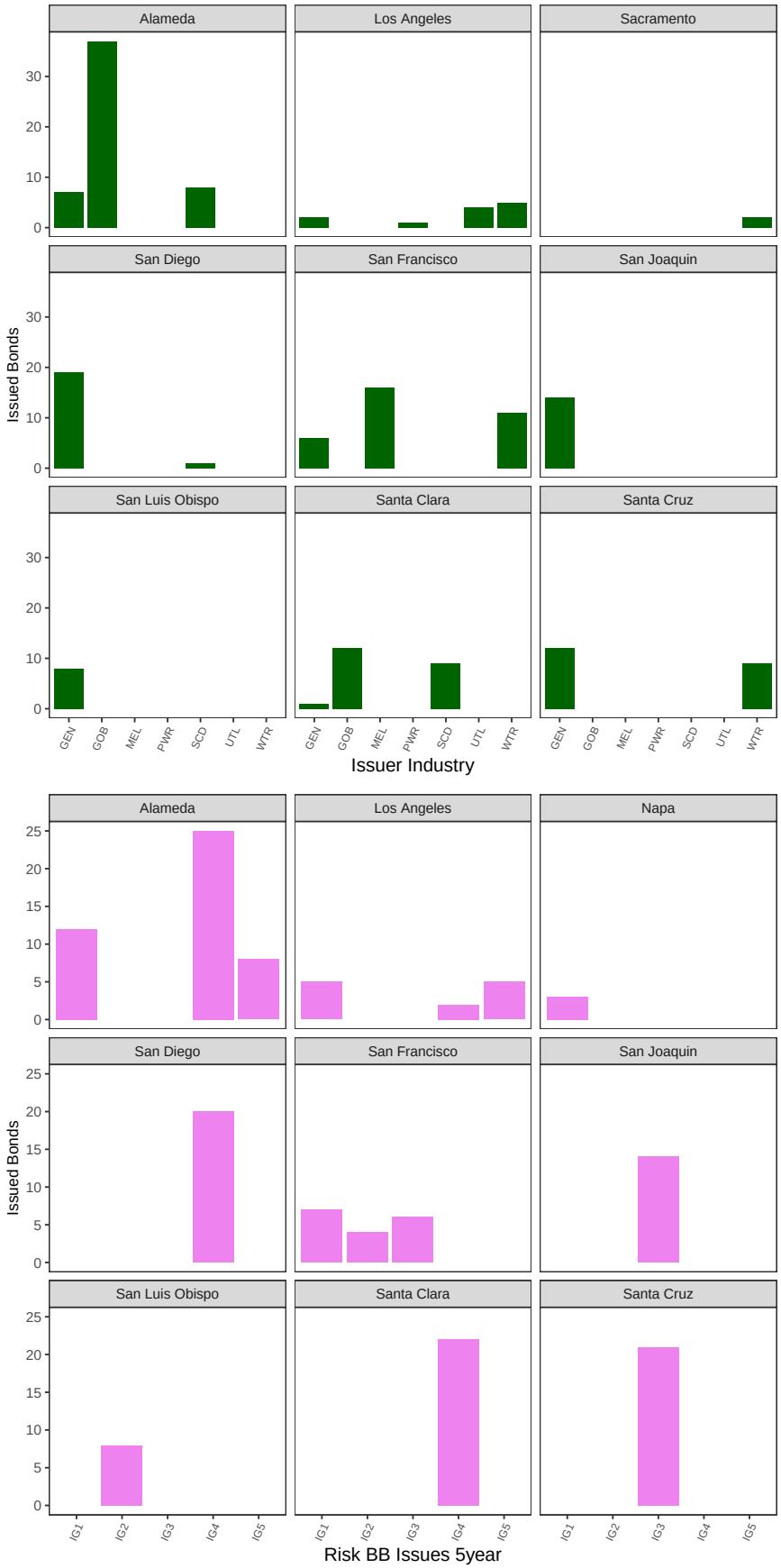
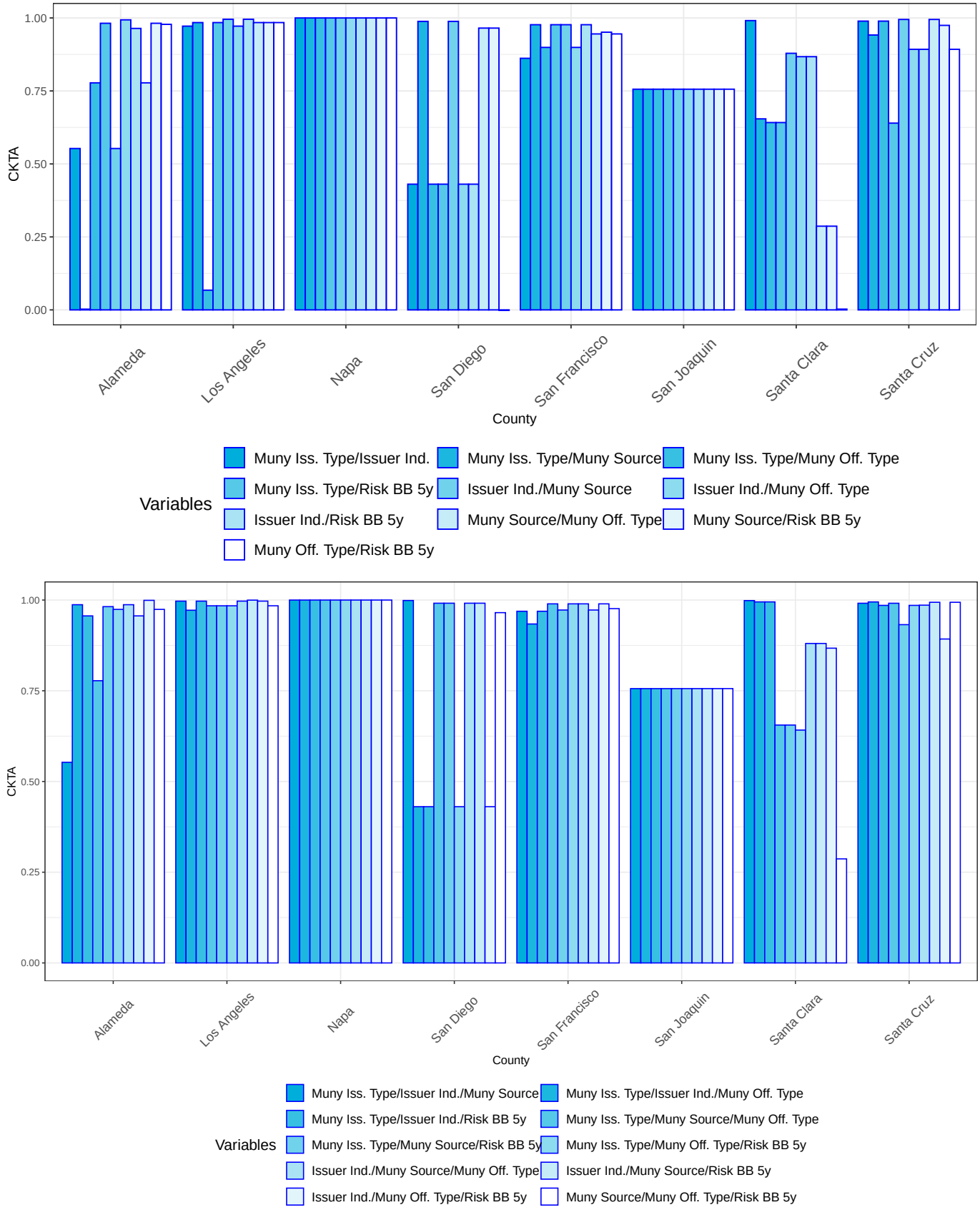


Figure 29: Barplots of some of the used financial variables split by county. The top plot refers to Muni Issue Type, followed by Muni Source, Issuer Industry and Bloomberg Issuer 5-Year Credit Risk.

E CKTA of Categorical Variables of the Financial Data set



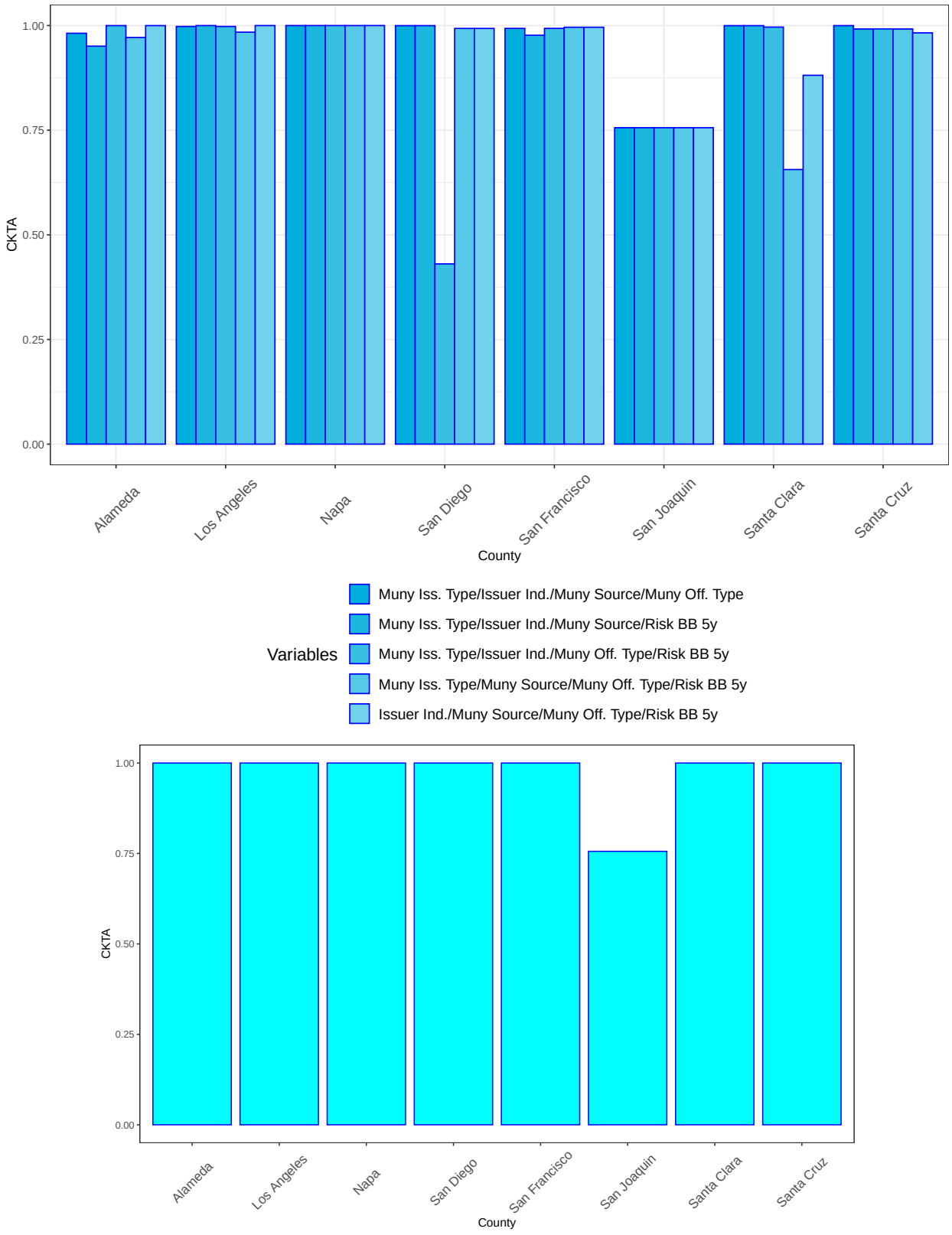


Figure 30: Relative contributions of each categorical variable to its original categorical data empirical covariance matrix. This is analysis is performed by county and by considering combinations 2 to 5 of the categorical variables. The cKTA is then computed by considering a combination of 2 (or 3,4 and 5) categorical variable to compute the first empirical covariance matrix and its equivalent one considering all of them, i.e. all 5 categorical variables, by county. The cKTA is comprised between -1 and 1.

F Qualitative Analysis PCs and kPCs

This Appendix presents the results of the qualitative analysis conducted on PCs and kPCs. The idea is to observe the behaviour of these base functions in capturing underlying data variations through smoothing splines. Fig. 31 refers to the PCs, while Fig. 32 to the kPCs.

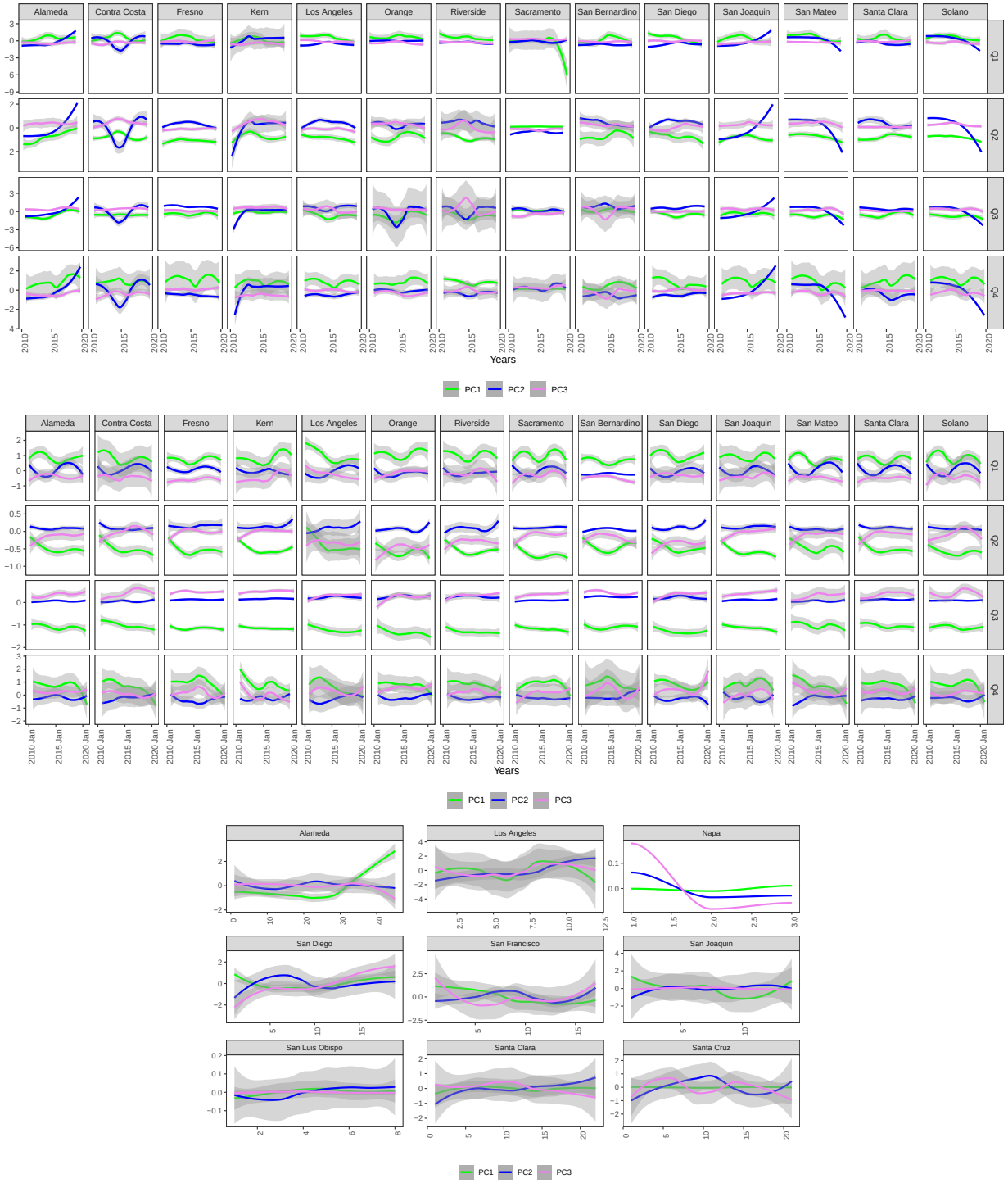


Figure 31: Results in PCA pollution, climate, and financial data sets. For visualisation purposes, we fitted smoothing splines through each PCs. Note that for the case of pollution and climate, we chose to observe the PCs by quarters.

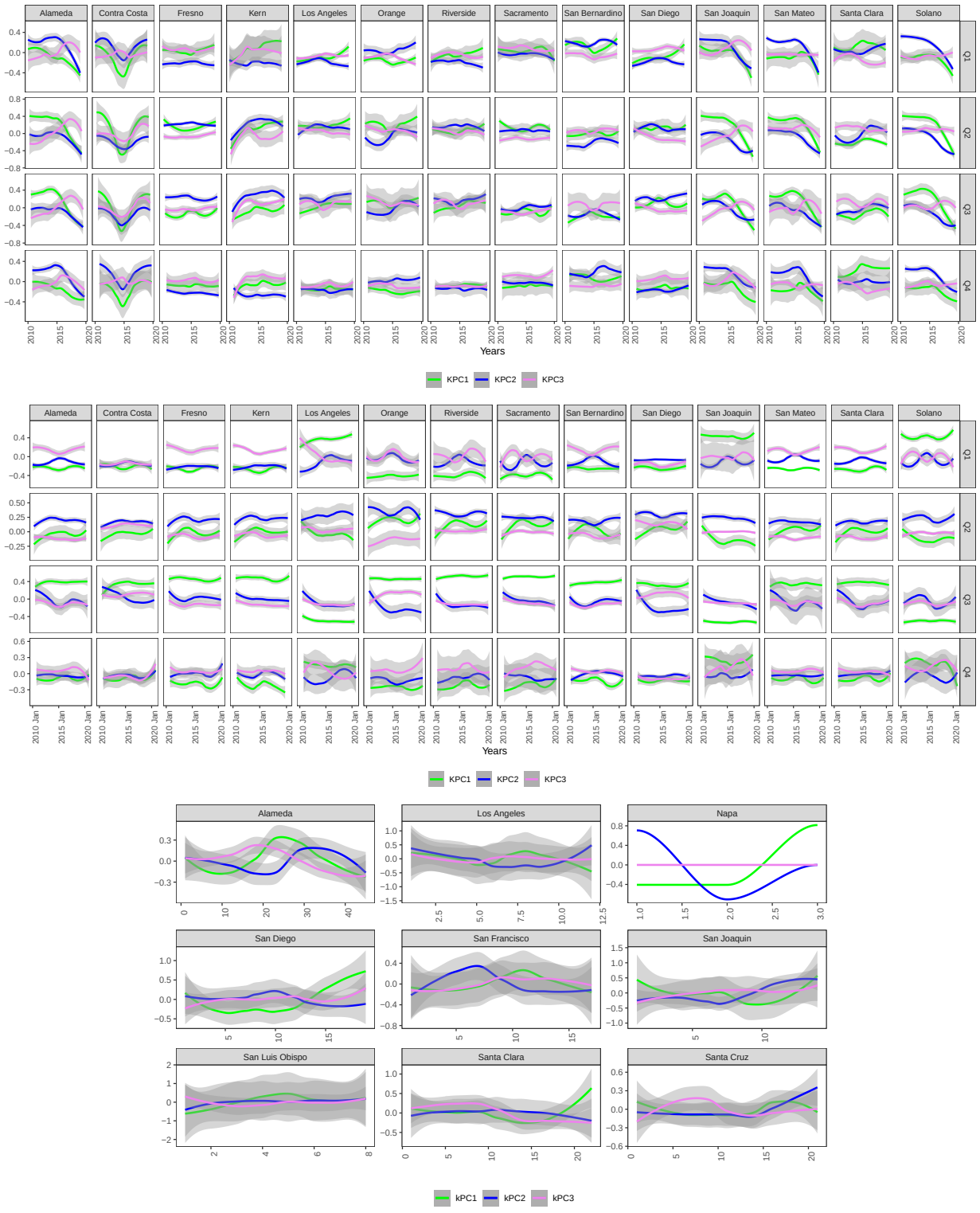


Figure 32: Results in kPCA pollution, climate, and financial data sets. For visualisation purposes, we fitted smoothing splines through each PCs. Note that for the case of pollution and climate, we chose to observe the PCs by quarters.

G Quantitative Analysis PCs and kPCs

This Appendix presents the results of the quantitative analysis conducted on PCs and kPCs. The idea is to observe the behaviour of these bases functions in capturing underlying data variations through boxplots of each basis by county and, for pollution and climate, by quarters. Fig. 31 refers to the PCs, while Fig. 32 to the kPCs.

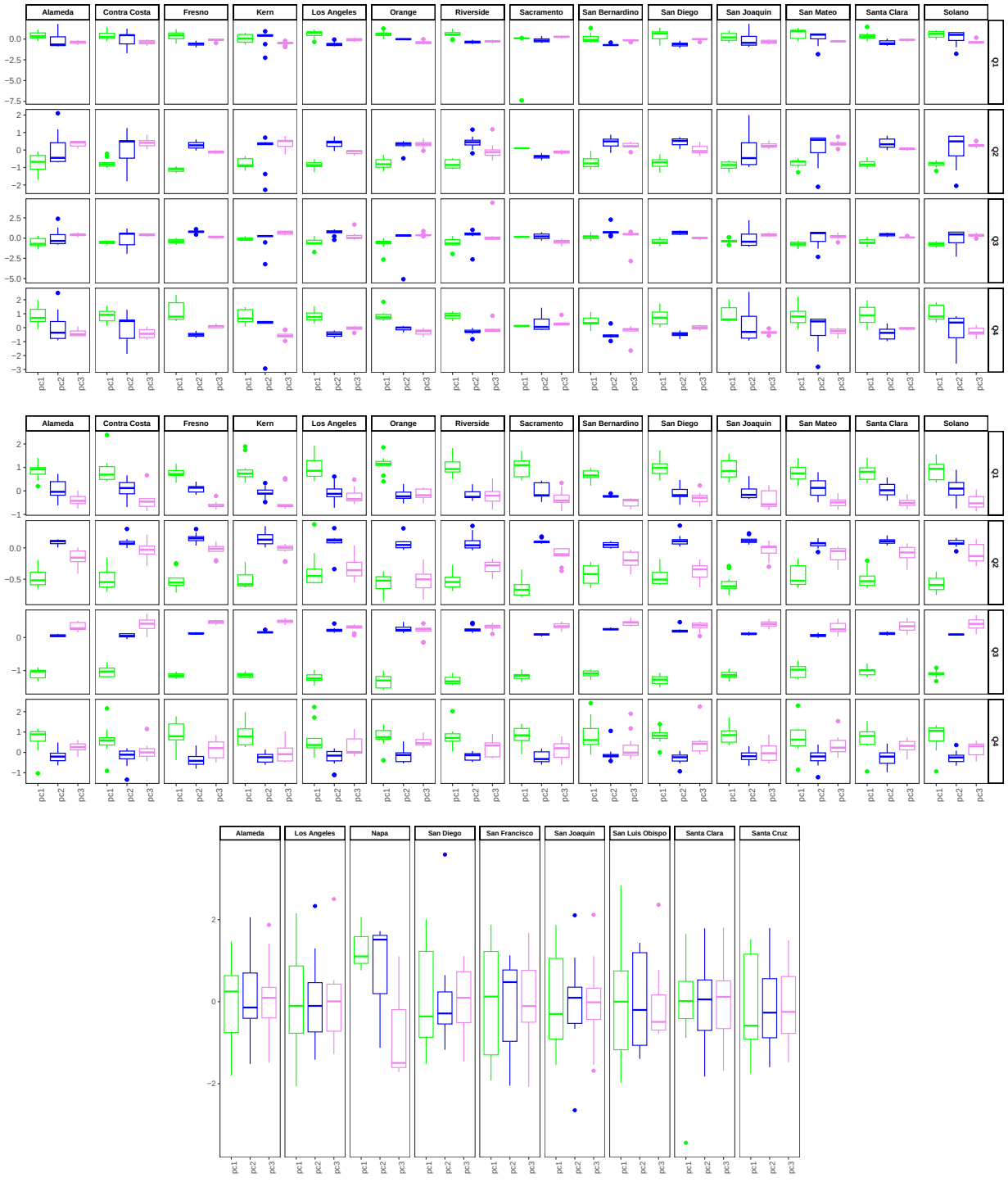


Figure 33: Results in PCA pollution and climate dataset and financial dataset.

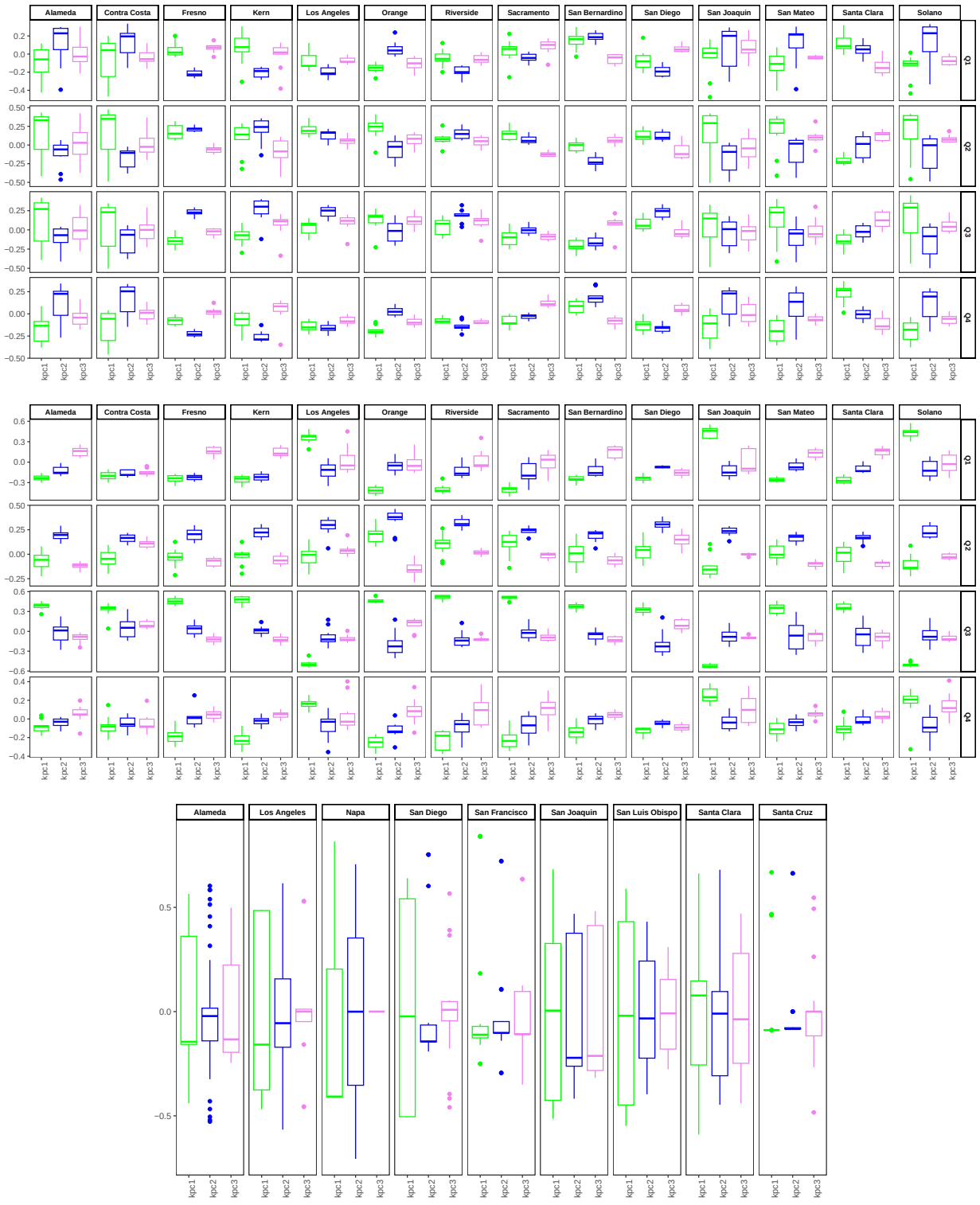


Figure 34: Results in kPCA pollution and climate dataset and financial dataset.

H Complete List of Tables (All the Counties)

Optimal γ Hyperparameter for The Datasets			
County	**Financial**	**Pollution**	**Climate**
Alameda	0.5	0.5	5
Contra Costa	0.5	0.5	5
Fresno	1.0	1.0	5
Kern	0.5	0.5	5
Los Angeles	0.5	0.5	1
Madera	1.0	1.0	1
Orange	1.0	1.0	1
Riverside	1.0	1.0	1
Sacramento	5.0	5.0	1
San Bernardino	0.5	0.5	5
San Diego	0.5	0.5	5
San Joaquin	0.5	0.5	1
San Mateo	0.5	0.5	5
Santa Clara	0.5	0.5	5
Solano	0.5	0.5	1

Table 23: Table describing the optimal γ parameters of the RBF kernel. 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 the population number of the considered counties (note that this selection criterion information is explained in detail in section 4). The columns represent the three different datasets, i.e. financial dataset, the pollution dataset and the climate dataset.

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**
Contra Costa	0.015	0.141	0.143	0.440	0.651	**0.803**
Fresno	0.449	0.596	0.699	0.579	0.526	0.525
Kern	0.548	0.503	0.623	0.699	**0.700**	**0.703**
Los Angeles	0.370	0.452	0.465	0.324	**0.740**	**0.808**
Orange	0.455	0.436	0.622	0.251	**0.790**	**0.796**
Riverside	0.050	0.388	0.445	0.049	0.612	0.595
Sacramento	0.011	0.514	0.549	0.647	0.669	**0.708**
San Bernardino	0.072	0.411	0.459	0.660	**0.751**	**0.810**
San Diego	0.280	0.530	0.551	0.262	**0.730**	**0.849**
San Joaquin	0.670	0.499	0.544	0.555	0.681	**0.719**
San Mateo	0.656	0.624	0.695	0.568	0.553	0.511
Santa Clara	**0.766**	**0.799**	**0.806**	0.681	**0.714**	**0.886**
Solano	0.313	0.313	0.515	0.444	0.601	**0.705**

Table 24: Table describing the cKTA obtained for the pollution data set. Note that in this table, we present the cKTAs for all the considered counties rather than just the one that we used for the PCA-CCA and kPCA-CCA. 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 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.

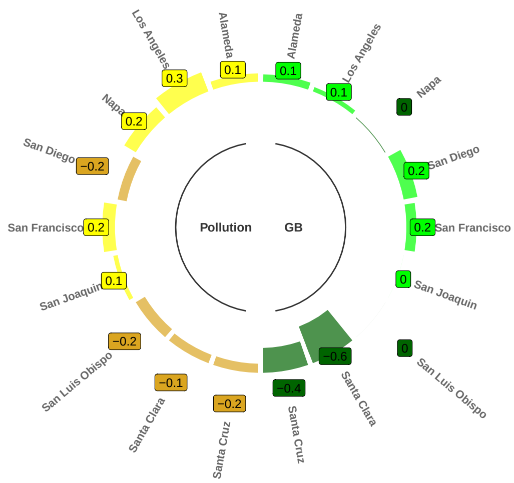
Results of Centered Kernel Target Alignment - Climate Dataset						
County	**PC1**	**PC1 & PC2**	**PC1 & PC2 & PC3**	**kPC1**	**kPC1 & kPC2**	**kPC1 & kPC2 & kPC3**
Alameda	0.434	**0.764**	**0.767**	0.515	**0.794**	**0.823**
Contra Costa	0.685	**0.799**	**0.842**	0.625	**0.887**	**0.953**
Fresno	**0.745**	**0.708**	**0.753**	0.621	**0.791**	**0.850**
Kern	**0.714**	0.652	**0.708**	0.458	0.617	0.666
Los Angeles	0.250	0.591	0.582	0.425	**0.733**	**0.842**
Madera	**0.771**	**0.736**	**0.771**	**0.703**	**0.735**	**0.841**
Marin	**0.734**	**0.821**	**0.821**	0.556	**0.794**	**0.840**
Napa	0.511	**0.818**	**0.841**	0.361	**0.819**	**0.920**
Orange	0.390	**0.718**	**0.715**	0.242	**0.823**	**0.836**
Riverside	0.605	**0.847**	**0.858**	**0.796**	**0.838**	**0.950**
Sacramento	0.469	**0.782**	**0.784**	**0.799**	**0.768**	**0.873**
San Bernardino	0.681	0.591	0.650	0.417	**0.726**	**0.897**
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 Mateo	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**
Solano	0.607	**0.832**	**0.853**	0.446	**0.787**	**0.803**
Stanislaus	**0.740**	**0.733**	**0.778**	0.382	**0.731**	**0.897**
Yolo	**0.714**	**0.866**	**0.866**	0.227	**0.854**	**0.931**

Table 25: Table describing the cKTA obtained for the climate data set. Note that in this table, we present the cKTAs for all the considered counties rather than just the one that we used for the PCA-CCA and kPCA-CCA. 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.

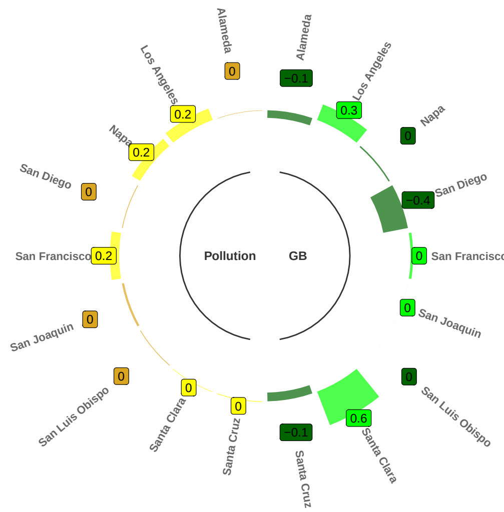
I More Canonical Variates Results

In this Appendix, we present further results for the structured coefficients of the second and third canonical variances of the kPCA-CCA. The following figures present results for both financial/pollution kPCs and financial/climate kPCs. Fig. 35 refers to the structured coefficients of kPCs of the pollution/financial data. The top panels present the results for the second and third canonical variates of kPC1, while the bottom panels for the second and third canonical kPC2.

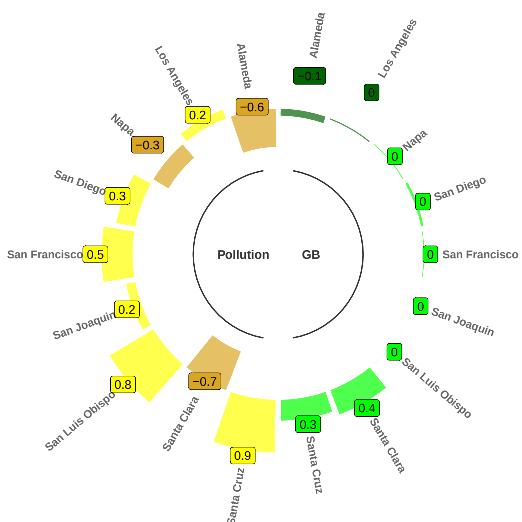
Fig. 36 refers to the structured coefficients of kPCs of the climate/financial data. The top panels present the results for the second and third canonical variates of kPC1, while the bottom panels for the second and third canonical kPC2.



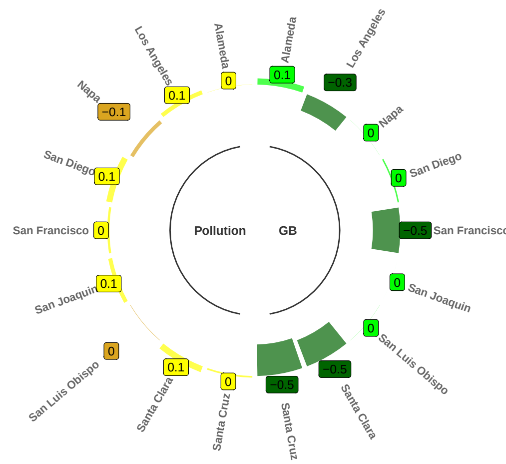
(a) kPC1-CV2.



(b) kPC1-CV3.

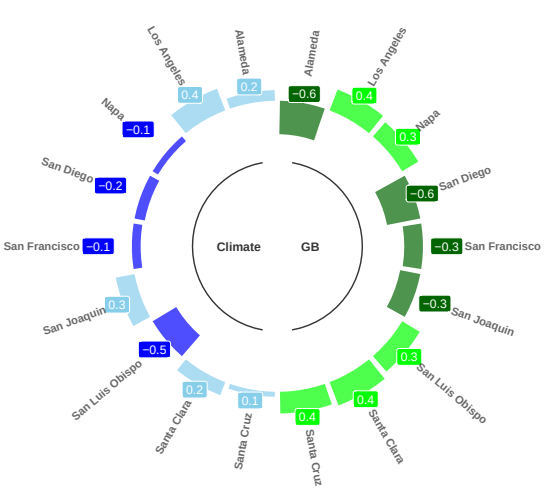


(c) kPC2-CV2.

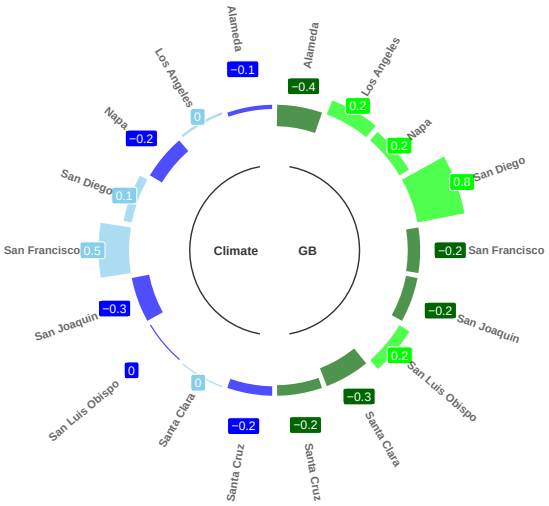


(d) kPC2-CV3.

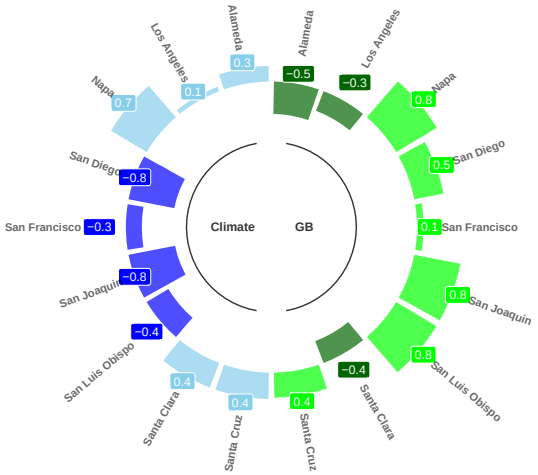
Figure 35: Structured coefficients of financial/pollution kPC1 (top) and kPC2 (bottom) for the second and the third canonical variates. Note that the results for the second canonical variate are presented in Table 21, in the fifth column.



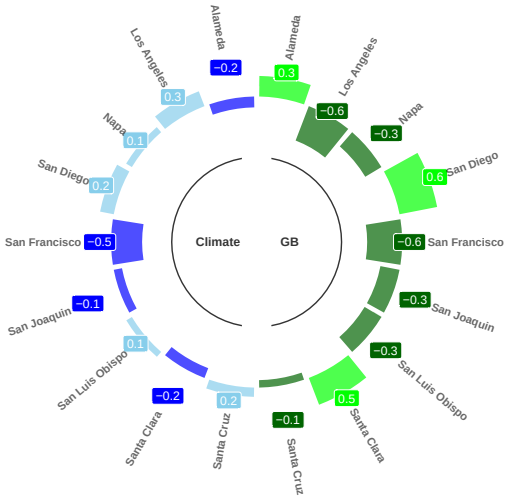
(a) kPC1-CV2.



(b) kPC1-CV3.



(c) kPC2-CV2.



(d) kPC2-CV3.

Figure 36: Structured coefficients of financial/climate kPC1 (top) and kPC2 (bottom) for the second and the third canonical variates. Note that the results for the second canonical variate are presented in Table 22, in the fifth column.