Smart Modelling and Large Data Sets

Cattani, Gilles

Smart Modelling and Large Data Sets

by

Gilles Cattani

A thesis submitted to the
Geneva School of Economics and Management,
University of Geneva, Switzerland,
in fulfillment of the requirements for the degree of
PhD in Econometrics

Members of the thesis committee:
Prof. Stefan Sperlich, Adviser, University of Geneva
Prof. Olivier Scaillet, University of Geneva
Prof. Miguel Delgado, Universidad Carlos III de Madrid

Thesis No. 116
November 2022

Abstract

In recent decades, semiparametric and nonparametric models have received increasing interest, which can be explained by the desire to get away from the strong restrictions of parametric models. Although their rate of convergence is slower, semiparametric and nonparametric models offer greater flexibility for estimation. This thesis proposes to use these models for respectively economic and econometric modelling in chapter one and two and to provide a solution to the distributed data problem in chapter three. In the first chapter, we use a general additive semiparametric model to estimate the long run efficiency of offshore wind farms. We rely on mainly well-established nonparametric methods that we had to modify appropriately to fit with the economic model we wanted to estimate. In the second chapter, we use a varying coefficient semiparametric model to estimate Regional Knowledge Production Function. We rely on general ideas of economic and econometric modelling of bilateral trade by gravity models and develop semiparametric estimators that could estimate such a sophisticated model structure. Finally in the third chapter, we start from the data distributed problem. We propose an operational way to get timely estimates, fits or predictions with huge but distributed data sets, including model and parameter selection. Our approach is to think fully locally using local linear nonparametric estimation with LASSO penalty for statistical analysis, may it be estimation, prediction, or attribution.

Résumé

Au cours des dernières décennies, les modèles semi-paramétriques et non-paramétriques ont fait l'objet d'un intérêt croissant qui peut s'expliquer par le désir de s'affranchir des fortes restrictions des modèles paramétriques. Bien que leur taux de convergence soit plus lent, les modèles semi-paramétriques et non-paramétriques offrent une plus grande flexibilité pour l'estimation. Cette thèse propose d'utiliser ces modèles pour la modélisation économique et économétrique dans les chapitres un et deux et d'apporter une solution au problème des données distribuées dans le chapitre trois. Dans le premier chapitre, nous utilisons un modèle semi-paramétrique additif général pour estimer l'efficacité à long terme des parcs éoliens en mer. Nous nous appuyons principalement sur des méthodes non-paramétriques bien établies que nous avons dû modifier de manière appropriée pour les adapter au modèle économique à estimer. Dans le deuxième chapitre, nous utilisons un modèle semi-paramétrique à coefficients variables pour estimer la fonction de production de connaissances régionales. Nous nous appuyons sur des idées générales de modélisation économique et économétrique du commerce bilatéral par des modèles de gravité et développons des estimateurs semi-paramétriques qui pourraient estimer une structure de modèle aussi sophistiquée. Enfin, dans la troisième partie, nous partons du problème de la distribution des données. Nous proposons une manière opérationnelle d'obtenir des estimations rapides, des extrapolations ou des prédictions avec de très grands ensembles de données distribuées, en incluant la sélection des modèles et des paramètres. Notre approche consiste à penser entièrement localement en utilisant l'estimation non-paramétrique linéaire locale avec une pénalité LASSO pour l'analyse statistique, qu'il s'agisse d'estimation, de prédiction ou d'attribution.

Contents

Introduction

In recent decades, semiparametric and nonparametric models have received increasing interest, which can be explained by the desire to get away from the strong restrictions of parametric models. Although their rate of convergence is slower, semiparametric and nonparametric models offer greater flexibility for estimation. This thesis proposes to use these models for respectively economic and econometric modelling in chapter one and two and to provide a solution to the distributed data problem in chapter three.

In the first chapter, we use a general additive semiparametric model to estimate the long run efficiency of offshore wind farms. We rely on mainly well-established nonparametric methods that we had to modify appropriately to fit with the economic model we wanted to estimate. Offshore wind energy has emerged as an attractive alternative to conventional resources to meet the Paris agreement commitment. This chapter studies the long run capacity of offshore wind farms to transform kinetic energy into electricity. We start estimating the technical efficiency of twenty-six farms over a thirteen years interval using a fully parametric and a semiparametric stochastic frontier model. The latter allows the factors of production to impact non-linearly on the quantity of electricity produced, those reducing the possibility of committing a functional misspecification error. Our results suggest that fully parametric specifications fails to identify the non-linear effect of labour cost on volumes of electricity produced. Then, we regress the estimated technical efficiency over the farm age, while controlling for the technological change of the wind power industry, to single out the resilience of the technical efficiency to aging. According to our calculations, technical efficiency ranges from 83% to 98% *and* it does not decline with age. This result shades light on the capacity of offshore wind farms to be a long term solution of the energy transition.

In the second chapter, we use a varying coefficient semiparametric model to estimate Regional Knowledge Production Function (RKPF). We rely on general ideas of economic and econometric modelling of bilateral trade by gravity models and develop semiparametric estimators that could estimate such a sophisticated model structure. The estimation of RKPF is subject of a vast literature prevailing the application of the spatial linear regression models. However, the adequacy of these models has been questioned in recent work which discloses the existence of nonlinearities and heterogeneous effects not sufficiently addressed in the existing literature on model specification. This chapter approaches these modelling issues recurring to some semiparametric methods that today are easily accessible to practitioners. We illustrate this along the analysis of panel data on European regional knowledge production. It is shown how the heterogeneity of effects of potentially complex functional forms in the RKPF can be revealed, including heterogeneous spatial spillovers. Among other alternatives, this work introduces varying coefficient spatial regression models for the RKPF, in which direct effects and spatial spillovers on knowledge creation due to variations of R&D expenditures and Human Capital resources depend on the population density of the region. Not surprisingly, we find a lot of heterogeneity in the effects and spillovers, creating a lot of serious non-linearities. Results obtained in the empirical study suggest for instance, that innovation policies should take into consideration the specific region features like population density.

In the third chapter, we start from the data distributed problem. We propose an operational way to get timely estimates, fits or predictions with huge but distributed data sets, including model and parameter selection. Our approach is to think fully locally using local linear nonparametric estimation with LASSO penalty for statistical analysis, may it be estimation, prediction, or attribution. We borrow ideas of local smoothers and prediction algorithms to generate our practical tool. Further, while typically distributed databases are considered as a bane, data localization can turn it into a boon. Similarly, since most of the problems with divide-and-conquer algorithms root in the paradigm of facing a global parameter set, they disappear by localization, and the selection of an optimal subsample size is melted with the one of optimal bandwidths which in addition we allow to be local too. Moreover, model and variable selection are possible, and sometimes even necessary, when staying local. For each step and subprocedure, we look for the most efficient implementation to keep the procedure fast. The proof of concept and computational details are given in a simulation study. An application to ocean warming illustrates the practical use of such a tool.

Chapter 1

Measuring the Long Run Technical Efficiency of Offshore Wind Farms

1.1 Introduction

During the last decade, offshore wind technology became a valuable alternative to conventional resources (Bosch, Staffell, & Hawkes, 2019). Measures like the generation per turbine and the generation per unit of capacity suggest that the opportunity-costs of deploying offshore hubs, which install bigger rotors and exploit faster and more uniformly distributed wind, is positive (Dismukes & Upton Jr, 2015). Furthermore, offshore wind resources seem to be a suitable option in different environments like the United States coast, the Iberian peninsula, the North Adriatic Sea and the South China Sea (Schweizer et al., 2016; Soares, Lima, Cardoso, Nascimento, & Semedo, 2017; Costoya, DeCastro, Carvalho, & Gómez-Gesteira, 2020; Wen, Kamranzad, & Lin, 2021). As a result, offshore technology could become a long term solution of the energy transition.

The previous conclusion holds true only if the capacity to generate offshore electricity does not decline over time. In other words, only if the Technical Efficiency (TE) of offshore farms remains constant for a number of years sufficient to amortize their deployment costs, the returns to scale can fully unfold. Previous studies suggest that this might not be the case. For example, regressing (ideal) load factors on the farm age suggests a statistically significant decline in onshore load factors over time (Hughes, 2012; Staffell & Green, 2014; Olauson, Edström, & Rydén, 2017). However, these conclusions seem to apply only to the onshore wind installations since an increase in offshore load factors have been observed between 2005 and 2014 in the United Kingdom (Crabtree, Zappalá, & Hogg, 2015). This increase could be explained either by an increase in offshore wind speed or a maturation of operation regimes, which reduces downtime. We proposed to investigate the effect of aging though TE analysis. To the best of our knowledge, no study has investigated the relation between the TE of offshore wind farms and their aging. The aim of the present paper is to close this gap in the renewable literature.

Applied econometricians estimate the TE using either Data Envelopment Analysis (DEA) or Stochastic Frontier Analysis (SFA). The former is a deterministic performance measurement, which assesses the relative efficiency of decision-making units (Charnes, Cooper, & Rhodes, 1978). The latter is a stochastic regression model, which separates the TE from random noises (Aigner, Lovell, & Schmidt, 1977; Meeusen & van Den Broeck, 1977). Traditionally, DEA has dominated the renewable literature due to the constraints SFA imposes on the structure of the production function (Akbari, Jones, & Treloar, 2020).

In recent years, SFA overcome many of its restrictions, while keeping the advantages of a probabilistic framework.

We use a flexible semiparametric SFA, which does not impose any strict output-input link, to identify the TE of the United Kingdom offshore wind industry. Our choice is motivated by the relative advantages of this methodology compared to traditional SFA and DEA approaches. On the one hand, the semiparametric SFA does not require a functional production function specification. On the other, the methodology remains stochastic. We start constructing a database, which merges information from the Renewable Energy Foundation, the United Kingdom companies register, and a wind speed estimation algorithm (Staffell & Pfenninger, 2016), to obtain the output and inputs of a standard neo-classical production function. The resulting panel data contains yearly information about twenty-six wind farms observed over the time interval 2005-2018. Then, we mimic the process introduced by Y. Fan, Li, and Weersink (1996), while adapting the first step to a semiparametric framework, to measure the farm-level TE.

Once obtained the efficiency measures, we regress them on the farm age, while controlling for technological change (Henningsen, 2014). Our empirical results suggest that no significant decline in efficiency is observed as time passes. This finding completes a series of previous results based on onshore installations (Iglesias, Castellanos, & Seijas, 2010; Barros & Antunes, 2011; Lin & Luan, 2020).

This study contributes in several aspects. Firstly, we propose a to our knowledge the first stochastic frontier analysis of offshore wind electricity production. Secondly, we decided to rely on a semiparametric extension of the traditional SFA to compute wind farm technical efficiency. Thirdly, we investigated if efficiency where affected by aging. Finally, we provided details on the construction of our database based on latest wind sector research.

The rest of the paper is organized as follows. Section 2 presents a semiparametric SFA applied to wind industry. Section 3 describes the construction of the dataset. Results are provided in section 4 and discussed in section 5. Section 6 presents the conclusions.

1.2 A Stochastic Frontier Analysis of the Wind Industry

Stochastic Frontier Analysis (SFA) is an econometric technique designed to identify the Technical Efficiency (TE) of a producing unit. At its core, there is the idea to disentangle the unexplained part of the production function into the *first error component*, which is a pure stochastic noise, and the *second error component*, which captures technical inefficiency.

The first SFA analysis applied only to standard production and cost functions (Aigner et al., 1977; Meeusen & van Den Broeck, 1977; W. H. Greene, 1980; Stevenson, 1980). While theoretically appealing, these rigid formulations could misspecify the input-output relation, those returning biased estimates of the TE (Giannakas, Tran, & Tzouvelekas, 2003). In recent years, Y. Fan et al. (1996) extended SFA to semiparametric specifications. These new formulations allow the different factors of production to impact non-linearly on the volumes of output produced, those decreasing the possibility to commit a functional form mispecification error. Among them, the present paper uses an additive semiparametric model where the outcome of the production function is the quantity of electricity produced and the inputs are the quantity of capital, labour and kinetic energy (Iglesias et al., 2010; Ferrara & Vidoli, 2017). In our framework, the logarithm of the quantity of electricity generated by offshore wind farm i at time t, measured in Megawatts Hour, is function of the logarithm of the quantity of capital employed, $k_{it} = \log K_{it}$, measured in thousands of pounds spent per year, of the cost of labour, $l_{it} = \log L_{it}$, measured in thousands pounds spent per year, and of the logarithm of the quantity of kinetic energy employed in the process, $e_{it} = \log E_{it}$, measured in Megawatts Hour,

$$\log(\text{Electricity})_{it} = \alpha + f_k(k_{it}) + f_l(l_{it}) + f_e(e_{it}) + \mu_i + \epsilon_{it} \ , \quad \epsilon_{it} = v_{it} - u_{it} \ . \tag{1.1}$$

In equation (1.1), α is the unconditional expectation of the produced electricity[1], $[f_k(.), f_l(.), f_e(.)]$ are three unknown smooth functions (Hastie, 2017), which can take any shape suggested by the data with one-dimensional convergence rates (Stone, 1986), and μ_i is a farm-specific fixed effect (i.e a specific intercept is estimated for each farm), which captures unobserved cross-sectional heterogeneity. The composite error ϵ_{it} is made out of two components: 1) the pure random noise $v_{it} \overset{iid}{\sim} \mathcal{N}(0, \sigma_v^2)$, which is normally distributed with homoskedastic variance, and 2) the technical inefficiency $u_{it} \overset{iid}{\sim} \mathcal{N}^+(0, \sigma_u^2)$, which is half-normal distributed with homoskedastic variance; see Weinstein (1964) and Aigner et al. (1977) for details.

We estimate the TE embedded in equation (1.1) employing the two-step procedure introduced by Y. Fan et al. (1996). First, we estimate equation (1.1) using a General Additive Model (GAM). From this first estimation, we extract the estimated residuals $\hat{\epsilon}_{it}$. Then, we estimate the ratio of relative variability of the two error sources, $\lambda = \frac{\sigma_u}{\sigma_v}$, maximizing the pseudo-likelihood function[2] ,

$$\max_{\lambda \in I\!R^+} \left\{ -n \log \hat{\sigma} + \sum_{i=1}^{n} \log \left[1 - \Phi(\hat{\epsilon}_{it} \hat{\sigma}^{-1} \lambda) \right] - \frac{1}{2\hat{\sigma}^2} \sum_{i=1}^{n} \hat{\epsilon}_{it} \right\} , \tag{1.2}$$

where $\Phi(.)$ is the cumulative density function of a standardized normal distribution and

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} \hat{\epsilon}_{it}^2 = \widehat{\sigma_u^2 + \sigma_v^2} \ . \tag{1.3}$$

Knowing $\hat{\lambda}$ and $\hat{\sigma}$, we obtain $\hat{\sigma}_u$ and $\hat{\sigma}_v$ solving a linear system of two equations in two unknowns. Finally, we write the probability density function of u_{it}, conditional on the estimated residuals $\hat{\epsilon}_{it}$,

$$f(u_{it}|\hat{\epsilon}_{it}) = \frac{1}{\sqrt{2\pi}\hat{\sigma}_*} * \frac{\exp\left\{ -\frac{(u_{it} - \hat{\psi}_*)^2}{2\hat{\sigma}_*} \right\}}{1 - \Phi\left(-\frac{\hat{\psi}_*}{\hat{\sigma}_*} \right)} \ , \quad \text{with} \quad \hat{\psi}_* = -\frac{\hat{\sigma}_u^2 \hat{\epsilon}_{it}}{\hat{\sigma}^2} \quad \text{and} \quad \hat{\sigma}_*^2 = \frac{\hat{\sigma}_u^2 \hat{\sigma}_v^2}{\hat{\sigma}^2} \ , \tag{1.4}$$

as a function of estimated quantities (Jondrow, Lovell, Materov, & Schmidt, 1982). The exponential of the opposite of the first moment of this distribution returns the TE of each farm,

$$\widehat{\text{TE}}_{it} = \exp(-\mathbb{E}[u_{it}|\hat{\epsilon}_{it}]) = \exp\left(-\frac{\hat{\sigma}\hat{\lambda}}{1 + \hat{\lambda}^2} \left[\frac{\phi(\frac{\hat{\epsilon}_{it}\hat{\lambda}}{\hat{\sigma}})}{1 - \Phi(\frac{\hat{\epsilon}_{it}\hat{\lambda}}{\hat{\sigma}})} - \frac{\hat{\epsilon}_{it}\hat{\lambda}}{\hat{\sigma}} \right] \right) , \tag{1.5}$$

where $\phi(.)$ is the probability density function of a standard normal distribution.

[1] $\alpha = E[\log(\text{Electricity})_{it}]$ is imposed for identification to avoid collinearity with fixed effects.

[2] Note that defining *lambda* as the ratio of variability allows us to maximize the log-likelihood function over a single parameter.

1.3 Dataset Construction

In order to fit equation (1.1), we collect information from three distinct data sources. First, we obtain annual farm level data on the volumes of Megawatt Hour produced downloading monthly data from the Renewable Energy Foundation website[3]. The monthly data are then aggregated to obtain annual ones. Second, we collect labour and capital expenditures using the United Kingdom companies register[4]. We start isolating the Special Propose Vehicles (SPV), which contain financial information specific to each farm. Annual reports are available at the register of United Kingdom companies. Then, we use the audited books of the SPVs, which are generally more reliable than information available from farm websites or public reports, to identify different types of expenditures (Ederer, 2015; Aldersey-Williams, Broadbent, & Strachan, 2019). More precisely, we use as a proxy variable for capital costs the Capital Expenditures (CAPEX), net of the decommissioning costs and of the values of the transmissions assets[5], which farms need to sell to the Offshore Transmission Network Owners (Aldersey-Williams et al., 2019),

$$k_{it} = \log(\text{CAPEX}_{it} - \text{Decommissioning CAPEX}_{it} - \text{OFTO Assets}_i) \ . \tag{1.6}$$

In the same way, we use as a proxy variable for labour costs the Operational Expenditures (OPEX) faced by the SPV. The latter includes sales and administrative costs,

$$l_{it} = \log(\text{Cost of Sales OPEX}_{it} + \text{Administrative OPEX}_{it}) \ . \tag{1.7}$$

Figure 1.1 shows that k_{it} and l_{it} are driven by the water depth and the distance to shore of the installation (Myhr, Bjerkseter, Ågotnes, & Nygaard, 2014). Since these two variables also explain a large fraction of the farm-specific capacity to generate electricity, it is reasonable to think that the combined effect of (k_{it}, l_{it}) and the farm fixed-effect μ_i, which isolates further farm-specific characteristics, like the type of turbine installed, leaves as the only elements of ϵ_{it} the technical inefficiency and a white noise.

Finally, we estimate the total wind energy flowing through the surface area of farm i at hour s using the formula of the kinetic energy,

$$e_{is} = \log\left(\frac{\rho}{2}\right) + \log A + 3\log w_{hjs} \ , \tag{1.8}$$

where $\rho = 1.23$ is the air density, A is the swept area of rotor, w_{hjs} is the wind speed at height h in location j, with $i \in j$, at a given hour s. While ρ and A are known quantities, vw_{hjs} must be estimated. Traditionally, this quantity has been computed for any h and j using annual averages (Iglesias et al., 2010). In order to improve the accuracy of our calculations, we use hourly estimates of wind speed at point j and height h relying on the algorithm proposed by Staffell and Pfenninger (2016). The authors record wind speed data from the Modern-Era Retrospective analysis for Research and Applications climate dataset[6]. This climate dataset maps the globe using a 0.5° × 0.66° grid, which reports

[3] The raw data can be downloaded at https://www.ref.org.uk/energy-data

[4] The raw data can be downloaded at https://www.gov.uk/government/organisations/companies-house

[5] All values are available on audited books.

[6] The row data are downloadable here https://gmao.gsfc.nasa.gov/reanalysis/MERRA/ and here https://gmao.gsfc.nasa.gov/reanalysis/MERRA-2/

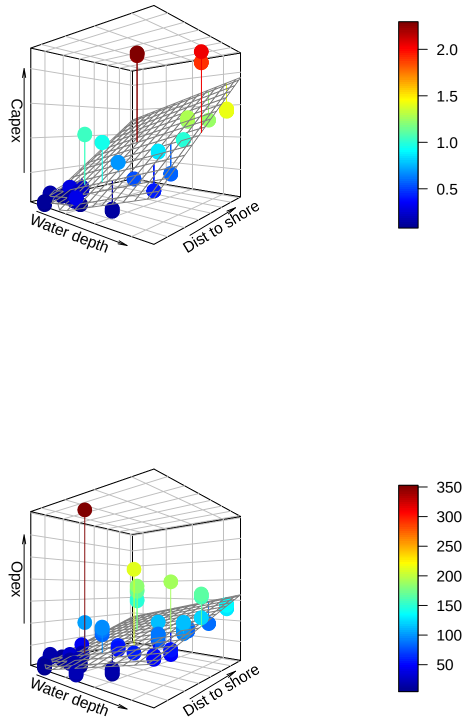


Figure 1.1: Estimated surface of CAPEX (billions pounds) and OPEX (millions pounds) against distance to shore (km) and water depth (meters).

hourly wind speed at every edge of the grid for three different heights. These initial data are obtained using the *climate reanalysis method*, a widely used meteorological technique, which combines historical observations and updates models in order to make retrospective forecasts (Dee et al., 2014). Once these initial data are collected, the algorithm interpolates hourly wind speed, at the desired location, using Locally Weighted Scatterplot Smoothing (LOWESS). Then, the algorithm extrapolates the hourly wind speed, at the desired height, using the logarithm Wind Profile Law (WPL) (Bañuelos-Ruedas, Camacho, & Rios-Marcuello, 2011),

$$w_{hjs} = \left(\frac{FV_{hjs}}{0.4}\right) \log\left(\frac{h - DH}{SR_{hjs}}\right) \ , \tag{1.9}$$

where FV is the friction velocity, 0.4 is the Von Karman constant, DH is the displacement height and SR is a measure of the surface roughness. Since the data contain measurements at three different heights, it is possible to estimate the unknown quantities (FV_{hjs}, SR_{hjs}) using a linear regression and obtain $\hat{w}_{hjs}$ for every height h, location j and hour s. Substituting the estimated $\hat{w}_{hjs}$[7] into equation (1.8), we compute the hourly kinetic energy of each farm. Then, we aggregate these estimates to obtain the annual kinetic energy of

[7] To take into account the limits of turbine technology, the author use only wind speed ranging from 0 to 40 m/s. In our case, all estimated wind speeds are ranging from 1.762 to 27.513 m/s.

each farm,

$$\hat{e}_{it} = \sum_{s=1}^{S} \left[\log \left(\frac{\rho}{2} \right) + \log A + 3 \log \hat{w}_{hjs} \right] , \tag{1.10}$$

where $S = 8,760$ during a normal year and $S = 8,784$ during a leap year. According to the Betz's law, wind turbines cannot transform more than 16/27 (59.3%) of the wind kinetic energy into electricity (Grogg, 2005). In addition to this physical barrier, there are aerodynamic, mechanical, technical and rated power limit losses, which contribute to decrease the ratio of kinetic energy transformed into electricity (Hau, 2013). We compute these ratios for each farm and each year based on our estimates and displayed resulting boxplots in Figure 1.2. The differences between the red dashed line representing the maximum possible energy conversion and boxplots correspond to wind farms energy conversion losses. Note that this figure highlights some heterogeneity due to farm specific fixed-variables, for instance turbine characteristics.

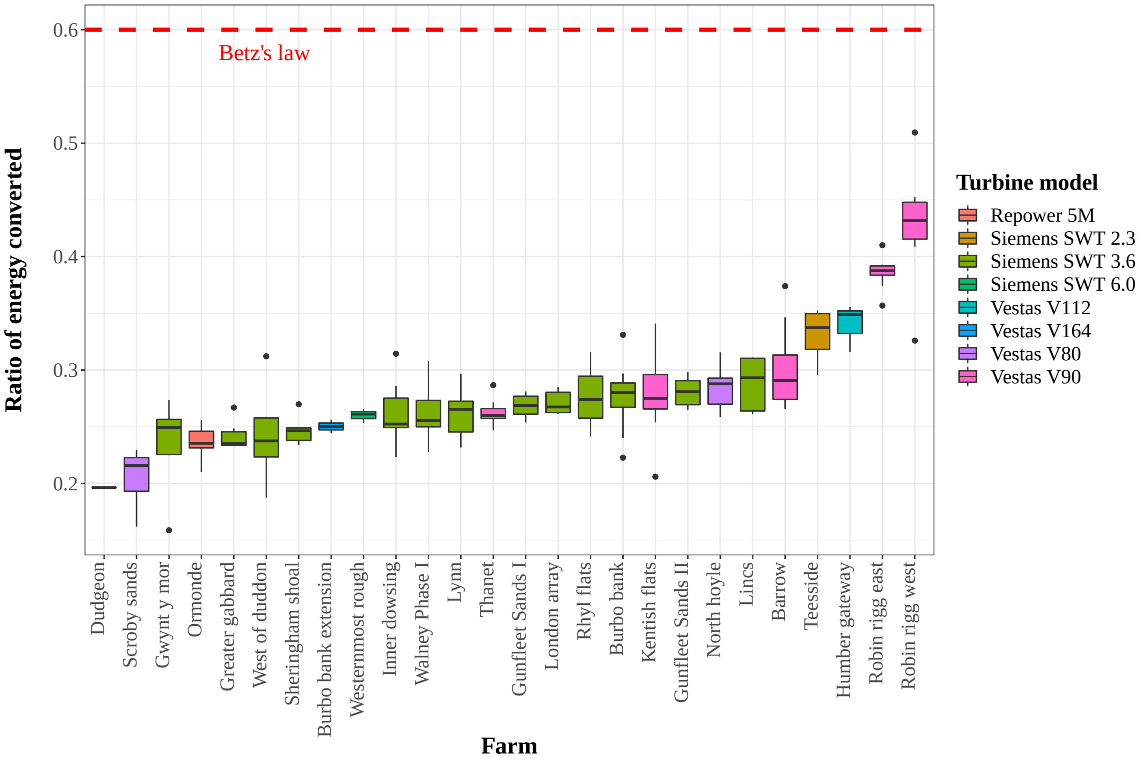


Figure 1.2: Ratio of Kinetic Energy converted into Electricity

Merging the information about electricity, costs and kinetic energy, we obtain a panel dataset, which contains yearly statistics for twenty-six offshore wind farms located in the United Kingdom across the time interval 2005-2018. The locations of wind farms are shown in Figure 1.3. In order to avoid transitional dynamics, we subtract from this initial dataset all the information regarding the year of the commissioning of the farm. The final dataset contains 176 data points[8]. The summary statistics of input variables of our production function defined in equation (1.1) are provided in Table 1.1.

[8] Note that a balanced dataset would have 26 * 14 = 364 data points. The discrepancy is due to the construction of several of the wind farms during the analyzed time interval.

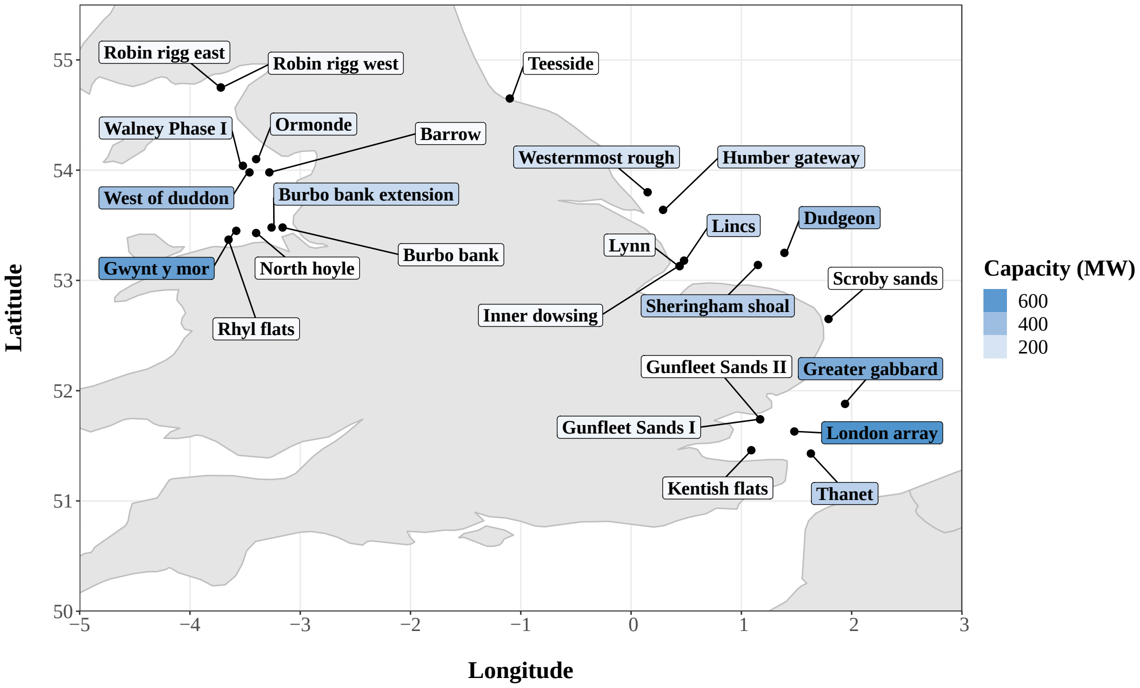


Figure 1.3: Location of Offshore Wind Farms

Table 1.1: Summary Statistics

Statistic	Variable	Unit of Account	N	Mean	SD	Min	Max
Electricity	Electricity	MWh	176	584,597	556,050	128,400	2,490,000
OPEX	L	1000 \$	176	46,512	53,139	4,561	352,802
CAPEX	K	1000 \$	176	477,237	545,687	77,561	2,296,264
Kinetic Energy	E	MWh	176	2,138,676	2,195,970	451,910	9,486,701

1.4 Empirical Results

We start our empirical analysis estimating a parametric version of equation (1.1),

$$\log(\text{Electricity})_{it} = \alpha + \beta_k k_{it} + \beta_l l_{it} + \beta_e \hat{e}_{it} + \mu_i + \epsilon_{it} \ , \tag{1.11}$$

which assumes that the three unknown smooth functions are linear, $f_k(.) = \beta_k k_{it}$, $f_l(.) = \beta_l l_{it}$, and $f_e(.) = \beta_e \hat{e}_{it}$. We estimate two versions of equation (1.11). In the first one, we do not take advantage of the panel nature of the dataset ($\mu_i \equiv 0$). In the second one, we control for the presence of unobserved fixed effects μ_i. This second specification corresponds to the true fixed effect model presented by W. Greene (2005). Then, we estimate equation (1.1), without imposing restrictions on the shapes of $[f_k(.), f_l(.), f_e(.)]$, using the thin plate regression splines option of the `mgcv` package of the statistical software `R` (Duchon, 1977; Wood & Wood, 2015). Like in the parametric case, we estimate a specification without and one with fixed effects. The obtained parametric terms of all four models are reported in Table 1.2.

In the parametric SFA without fixed effects, all factors of production impact positively on the volumes of electricity produced. The most important one is the amount of kinetic energy (i.e. *the fuel* of the farm). Increasing the kinetic energy by 1% increases the electric output by 0.62%. The economic variables are also relevant. Increasing CAPEX by 1% increases the electricity production by 0.12%, increasing OPEX by 1% increases the electricity production by 0.15%. Once the fixed effects are introduced, both economic variables become statistically insignificant. This result suggests that if we control for unobserved heterogeneity the only relevant factor is the quantity of kinetic energy, those there is no statistically significant degree of substitutability between money, allocated either in operational either in capital expenditures, and wind. Finally, both parametric models indicate decreasing returns to scale since the sum of $(\hat{\beta}_k, \hat{\beta}_l, \hat{\beta}_e)$ is smaller than one in both cases.

Like in the case of the parametric model, in the semiparametric SFA without fixed effects, all factors of production are statistically significant. However, contrary to the parametric model, in the semiparametric SFA with fixed effects, OPEX remains statistically significant. To investigate this difference, we display in Figure 1.4 $\hat{f}_l(.)$ and $\hat{f}_e(.)$. The former is non-linear and its impact changes in sign as OPEX grows. This finding is overlooked by the parametric specification. In other words, increasing OPEX can increase the farm output at least till a neighborrod around 60,000 pounds spent. To the contrary, like in the parametric case, any increase in kinetic energy always augments the final output.

Table 1.2: Estimated Stochastic Frontier Models

	Parametric SFA		semiparametric SFA	
	(1)	(2)	(3)	(4)
Intercept	0.959*** (0.221)		12.925*** (0.008)	
k	0.125** (0.040)	0.150 (0.113)		
l	0.149*** (0.037)	0.005 (0.030)		
e	0.624*** (0.032)	0.406*** (0.064)		
σ	0.139	0.091	0.121	0.073
λ	0.000	1.208	1.076	0.957
Fixed Effect	No	Yes	No	Yes
Observations	176	176	176	176
AIC	-185.51	-366.76	-272.74	-405.19

Note: *p<0.05; **p<0.01; ***p<0.001

Following the procedure presented in Section 2, we use $(\hat{\sigma}, \hat{\lambda}, \hat{\epsilon}_{it})$ to obtain $\widehat{\text{TE}}_{it}$. According to our estimates, technical efficiency is relatively high both in parametric and in semiparametric fixed effects specification with values ranging from 0.835 to 0.986, with the semiparametric estimates showing a smaller variance. Figure 1.5 displays the emipirical probability density function of $\widehat{\text{TE}}_{it}$.

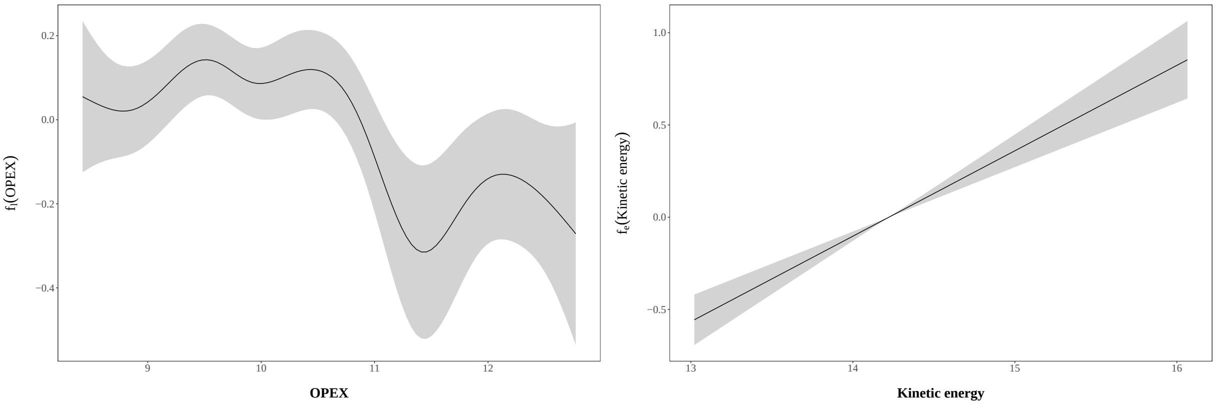


Figure 1.4: Estimated Smooth Functions $\hat{f}_l(.)$ and $\hat{f}_e(.)$.

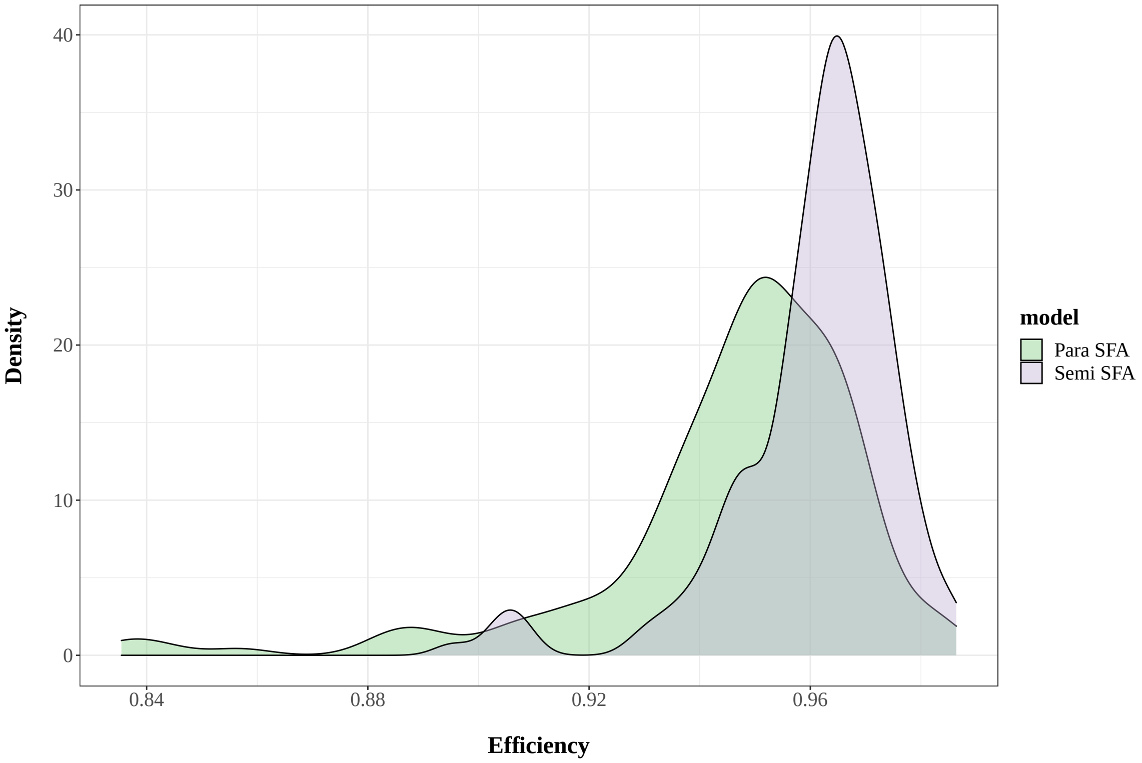


Figure 1.5: Estimated Empirical Probability Density Function of $\widehat{\mathrm{TE}}_{it}$.

Once obtained the technical efficiency, we can investigate its origin regressing $\widehat{\mathrm{TE}}_{it}$ on a set of explanatory variables. Our key interest is to understand if this measure is impacted by the aging of the farm. Therefore, we regress it on the farm age, while controlling for a time trend (i.e corresponds to the year of each observation),

$$\widehat{\mathrm{TE}}_{it} = \underset{(0.9072)}{0.4520} + \underset{(0.0004)}{0.0001} \ \mathrm{Age}_{it} + \underset{(0.0005)}{0.0003} \ \mathrm{Year}_t \ , \tag{1.12}$$

which should incorporate an eventual homogeneous technological trend within the wind industry. The latter should increase the technical efficiency at rate $\partial\widehat{\mathrm{TE}}_{it}/\partial\mathrm{Year}_t$. According to our result, both Age_{it} and Year_t do not impact in a statistically significant way $\widehat{\mathrm{TE}}_{it}$, see Figure 1.6. We confirm that conclusion using bootstrap estimated standard

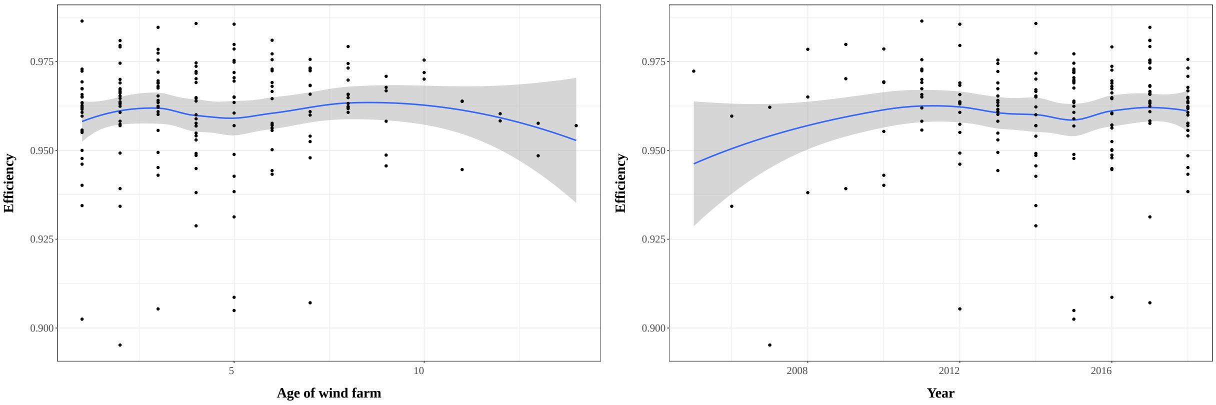


Figure 1.6: Offshore wind farms efficiencies against age and year.

errors. We also check if variables Age_{it} and Year_t are jointly significant using an F-test. The test reject the joint significance of the variables. Furthermore, the regression has no explanatory power. This second result confirms our distributional assumption of technical inefficiency, $u_{it} \overset{iid}{\sim} \mathcal{N}^+(0, \sigma_u^2)$.

1.5 Discussion

Our results contradict a series of previous measures of onshore wind farm performance decline with age based on linear regression. Hughes (2012) measured a decline in wind farm capacity factor of 5-13%[9] per year in the United Kingdom and in Denmark. A more consistent estimation of 1.6% decrease per year is obtained by Staffell and Green (2014) using corrected capacity factors of onshore wind farms located in United Kingdom. Recently, a similar study has obtain a significantly lower estimated decline of 2.7-5.4% decrease over 20 years in Sweden onshore wind farms (Olauson et al., 2017). Two explanations may explain why our results are contradictory to the current literature. First, it is possible that new offshore turbines require more than ten years to display a statically significant decline in their performances (Astolfi, Byrne, & Castellani, 2021). Although our panel includes a time period of thirteen years, there are relatively few wind farms older than ten years, as shown in the figure 1.6. Hence, it might be too soon to observe a decrease in offshore wind farms efficiency. This finding would be per se interesting because it would suggest that there is no decline in TE for more than a decade from the installation of the mills, this might not be the case for onshore installations as a decrease in load factors has been observed over time (Hughes, 2012; Staffell & Green, 2014; Olauson et al., 2017). Second, in the cited studies *performance* was measured through the transformation of a single input, namely the kinetic energy. Whereas in our paper, efficiency is measured by a multivariate production frontier made out of three inputs. A decline in the load factors could then be compensated by a more efficient use of capital and labour. Said differently, there would be a statistically significant substitution effect between natural capital (i.e. the wind), physical capital (i.e. CAPEX), and human capital (i.e. OPEX).

[9] These results are controversial as they are likely to be mainly driven by farm downtime

1.6 Conclusions

To the best of our knowledge, no empirical study has investigated the effect of aging on the TE of offshore wind farms. The present paper tries to fill using data of the United Kingdom offshore wind sector.

Starting from a standard production function, which links the quantity of electricity produced to the capital, the labour and the kinetic energy used, we estimate a parametric and a semiparametric stochastic frontier analysis. We obtain the TE of twenty-six farms observed over a thirteen years period. According to our results, once netted out farm-specific fixed effects, the TE ranges from 83% to 98% and are not affected by aging of the farm. This result suggests that offshore wind farms are highly efficient in the short as well as in the long run.

This empirical result complements the literature on wind power across three aspects. First, we show the first application of a Stochastic Frontier Analysis (SFA) to the offshore wind sector. Second, among different types of SFA, we apply a flexible semiparametric formula, which relaxes many of the traditional SFA assumptions and captures the non-linear impact of operational expenditures on the quantity of electricity produced. Third, we describe in detailed how to construct the database to allow other researchers to replicate this type of research for different regions of the world.

Chapter 2

Nonlinearities and Heterogenous Effects in the Regional Knowledge Production Function

2.1 Introduction

It has broadly been recognized that spatial econometric methods are useful tools to model an insightful data analysis, and have therefore become eminently requested with the increasing availability of geo-referenced data. Since the seminal work of Paelinck (1978) the specification modelling and estimation techniques have evolved profusely in this direction. Instead of a comprehensive revision, we refer to Anselin (2010) as a quite important survey on the evolution of spatial econometrics since its outset till the last decade. In particular, it displays well the increasing importance of the subject moving from the margins of applied regional science to the mainstream of econometric methodology. See also Arbia (2016) for a brief overview of recent developments in the area. In this context, already McMillen (2012) acknowledged the importance of using semiparametric and nonparametric techniques in model specification and estimation with spatial data. He emphasized its capacity to deal with the prevalent problems of correlated unobserved spatial heterogeneity and unknown functional form due to the complexity of spatial relationships, see also Pinkse and Slade (2010). More specifically then, Basile, Durbán, Mínguez, Montero, and Mur (2014) introduced a semiparametric spatial regression model based on penalized splines, whereas Lu, Steinskog, Tjøstheim, and Yao (2009) modelled spatial heterogeneity in covariates' impacts for spatio-temporal data proposing by coefficients varying over spatial location and an unknown index. They estimated both, i.e., the coefficient functions and the index variable with a two-step procedure that recurs to standard kernel regression. More recently, Basile and Mínguez (2018) provided a critical discussion comparing parametric and semiparametric spatial regression models.

Independently from the above said, the ability of regions to produce knowledge and innovation is an important subject of research. Griliches (1979) introduced the knowledge production function (KPF) as a tool to analyse the creation of knowledge and innovation. This approach has not lost anything of its popularity until today. It departs from the fact that knowledge is an output determined by R&D investment and human capital resources. Somewhat more recently, several authors disclosed that regional knowledge production has to take into account the effect of knowledge endowments of the neighbours in the transmission and generation of knowledge. As this obviously is - at least partly - reflected in the spatial proximity of regions it clearly suggest to recur to spatial econometrics which introduce explicitly spatial dependency and spatial effects in model specification and estimation. On the other hand, the complexity in the relationship between R&D expenditures and the creation of knowledge is documented in the literature, leading to refinements of the KPF, e.g., whether including additional variables that capture regional features like industrial diversity (Piribauer & Wanzenböck, 2016), technological proximity and intensity of economic activity (Parent & LeSage, 2008), the level of foreign direct investment (F. Zhang, Wang, & Liu, 2020), or considering more flexible specifications by introducing nonlinearities in the KPF (Piribauer & Wanzenböck, 2016; Kijek & Kijek, 2019), spatial heterogeneity in the effects of the input variables (Autant-Bernard & LeSage, 2019; Kang & Dall'erba, 2016; Parent & LeSage, 2008). The latter arguments call for flexible modelling, recurring to nonparametric and semiparametric approaches (Charlot, Crescenzi, & Musolesi, 2015) as we discussed them above. This link has been motivating the present paper.

This paper contributes to the literature in at least two ways. On one hand, it contributes methodologically by analysing the complex relations between knowledge creation and knowledge inputs using semiparametric spatial models and allowing for spatial spillovers. It gives a special relevance to spatial econometric models with varying coefficients, where the effects of each variable may depend on other variables, simultaneously allowing for spatial dependency and spatial effects. On the other hand, it contributes to a better understanding of how to model the complex relationship between R&D investment, Human capital and knowledge production, making use of methods that today are accessible to the user. The main objective is to explore and understand better the heterogeneity of effects; classical methods only estimate direct average effects. This is little informative - and thus not of much help - if the effects' heterogeneity is of first order, i.e. more important than the average effect itself. To our knowledge, those varying coefficients models (in the perspective that marginal effects of a variable may depend on other variables, often called drivers, see Sperlich and Theler (2015), (Benini, Sperlich, & Theler, 2016), Benini and Sperlich (2021)) are not used in spatial econometrics and in particular, cannot be found in the KPF literature. In a different context, using firm level data, Kanwar and Sperlich (2019) consider a varying coefficients conditional difference-in-differences specification to assess the impact of the intellectual property environment in India in factor productivity assuming that the first depends on the level of R&D activities of the firms.

It should be mentioned that the augmented KPF approach has been criticised, among others by Ó hUallacháin and Leslie (2007), which argue that the added regional features into the KPF are a source of confusion between causes and effects. Autant-Bernard and LeSage (2019) reinforce this idea and prefer to use instead a heterogeneous spatial autoregressive model for spatial-temporal data. Kang and Dall'erba (2016) present a comprehensive survey on the approaches used in the literature to model spatial heterogeneity in the KPF. However, very few consider spatial heterogeneous marginal effects. We distinguish the work of Autant-Bernard and LeSage (2019) with spatial-temporal data where all the coefficients are specific to a region (vary across regions), estimated by Bayesian methods and ridge regression to overcome collinearity problems. The latter emerge because in their estimation, they need a significant variation over time to identify the region-specific coefficients. In some regions, however, the variables hardly vary over time. Kang and Dall'erba (2016) use a different approach, recurring to geographically weighted regression (GWR) and mixed GWR in a sample of US counties. They observe a high degree of spatial heterogeneity in the marginal effects of the knowledge input variables across Metropolitan Statistical Areas. To the best of our knowledge, no-one has so far considered the different alternatives we discuss below, showing how an explorative spatio-temporal data analysis of a KPF could be performed in practice, resorting only to ready-to-use software.

In the next section we review the most related RKPF literature to place our contribution. Then we briefly introduce the data along which we have organized the presentation and discussion of the data analysis and modelling which we present in the section that follows. We conclude with a brief discussion, and have deferred some more tables and figures to the appendix.

2.2 Knowledge spillovers, nonlinearities and omitted heterogeneity in the RKPF

Since the seminal work of Griliches (1979) a profusion of empirical articles has been published on estimating Knowledge Production functions. A special interest has been given to the regional knowledge production embedding the existence of spatial knowledge externalities and potential presence of spatial dependence in the error terms of traditional regression models. These could also result from omitted determinants of innovation creation with high spatial heterogeneity (Autant-Bernard, 2012). For more discussion see also Audretsch (2003) for an analysis on the role of spatial spillovers and externalities in the production of knowledge. On the other hand, Autant-Bernard and LeSage (2019) stress that there are a variety of region-specific features such as social and business network structure, social and demographic stratification, educational achievement, cultural factors, governance, or science–industry relationships, that influence the creation of regional knowledge. The usual proxies used in the empirical literature for the inputs of knowledge production are unable to capture well the effects of all those regional features, resulting in specifications for the RKPF with omitted heterogeneity.

To address the presence of spatial dependence and knowledge spillovers in the RKPF, some spacial econometric tools have been used, typically following the pioneer work of Anselin, Varga, and Acs (1997). Firstly, traditional linear spatial regression models were widely adopted. A higher accuracy when estimating the complex relation between knowledge creation and its main inputs (R&D expenditures and Human Capital) was expected to be assured with a judiciously choice of the proxy control variables of RKPF. These aim to disclose effects of unobserved spatial heterogeneity due to spacial and technological proximity between regions, regional economic dynamics, institutional environment, among others. See, for example, Buesa, Heijs, and Baumert (2010) for a study on the choice of the determinants of regional knowledge creation, and (Ferreira & Godinho, 2015) for different proxies to control in particular for technological sophistication, regional development, entrepreneurship and institutional environment regulation. At the same time, the inclusion of spatial lags in the RKPF has become popular as a mean to control for omitted heterogeneity in the classical specification, and consequently to prevent from a potential endogeneity bias in estimation. Autant-Bernard and LeSage (2011) give an empirical motivation for this approach, stressing as well its advantage of enabling the identification of direct and indirects effects or spillovers in innovation activities. The estimation of spatial knowledge spillovers (or externalities) is an important issue in the empirical literature on innovation and knowledge creation; see, among others, Bottazzi and Peri (2003), Moreno, Paci, and Usai (2005b), Autant-Bernard and LeSage (2011) or Kijek and Kijek (2019). Parent and LeSage (2008) use a different approach. They estimate a Bayesian hierarchical model in which regional knowledge spillovers are captured by latent random coefficients that are spatially structured, accounting for the connectivity structures between regions by relying on technological as well as transportation and geographical proximity.

More recently, with the availability of panel data, or more generally, data with repeated observations of spatial units over several time periods, and the development of spatial panel methods, several authors opt to estimate the RKPF recurring to space-time models. As the name says, these models allow to analyse the dynamics of knowledge production over space and time as well. To this end, dynamic models are popular containing spatial lags of the variables together with time lags. Moreover, the inclusion of fixed or random effects specific to time and/or to the spatial unit allows to control for omitted heterogeneity in order to gain robustness against omitted variables (using fixed effects) or to increase efficiency (using random effects). As examples, see (Piribauer & Wanzenböck, 2016) who specify a linear dynamic space-time KPF for European regions, extending the classical linear spatial Durbin model with space and time lags, adding, as well, region and time fixed-effects to control for region-specific and time-specific omitted heterogeneity, and Parent (2012) who considers a linear spatial dynamic panel data model for knowledge creation in the US states with random effects, estimated by Bayesian Markov Chain Monte Carlo methods. Another motivation to include fixed/random effects in the modelling is to avoid spurious spatial dependence. Because the omitted variables causing heterogeneity might be spatially correlated. Therefore if there are not properly controlled, they could cause spurious spatial dependence (Heckman et al., 1981).

Autant-Bernard and LeSage (2019) advocate the existence of regional disparities in the ability to transform local R&D and Human capital inputs into innovation, and to benefit from or generate interregional spillovers. The parametric RKPF that is typically used in the empirical literature does not account for these regional heterogeneities even for space-time panel data models, because the coefficients associated to the inputs of knowledge creation are the same over all time periods and regions except if one would include several interaction terms (i.e., interacting production factors with space and time indicators). To overcome this limitation the above authors introduce a heterogeneous coefficient spatial autoregressive (HSAR) model that allows for variations in the level of spatial dependence/interaction as well as in the RKPF coefficients, intercepts and noise variances across each region. Their approach allows also to introduce prior information in their estimation (via Bayesian modelling). The estimator relies on Markov chain Monte Carlo (MCMC) procedures in place of a maximum likelihood or quasi-maximum likelihood (QML) based procedure.

The vast majority of empirical articles in the literature of regional knowledge creation estimate a log-linear RKPF. However, the linearity of the RKPF has repeatedly been questioned. Proença and Glórias (2021) argue the nonlinear functional form of the Cobb-Douglas type of the RKPF should be directly estimated through Poisson Quasi-maximum Likelihood because it could lead to more accurate estimates of the direct and indirect effects of knowledge inputs than those obtained with the loglinear model. On the other hand, (Griliches, 1990) alerts for the complexities in the process of knowledge creation leading to nonlinearities in the knowledge production functional form, though this issue has been largely neglected in the empirical literature until recently. Charlot et al. (2015) disclose empirically important nonlinearities in the RKPF by estimating a semiparametric Generalized Additive Model (GAM) with spatial effects to account for spatial dependence, including time and region specific fixed effects, recall our discussions from above. Basile and Mínguez (2018) in their above-mentioned critical review of parametric and semiparametric spatial econometric models specify a Penalized-Spline Spatial Lag model (with a spatial autoregressive component) that would account for spatial dependence and nonlinearities in the functional form. They control for unobserved spatial heterogeneity by including a geo-additive term which is a smooth function of the spatial coordinates of the regions.

The semiparametric approaches mentioned prove to be very flexible to model the complex process of knowledge creation. However, given those methods are intrinsically data-driven, they can easily produce results that are hard to interpret or even odd. Moreover, there is not only the risk of scarifying interpretability for flexibility. If not data fitting and prediction but interpretation of the model parameters and functions is the central interest, then this should be reflected in the modelling. Finally, even if one agrees on a flexible modelling for interpretation (instead of prediction and data-fitting), you can still distinguish between targeted and untargeted modelling. For example, using a random slope coefficient may reflect very well the heterogeneity of the associated covariate's effect, but it cannot 'explain' it. For this you may rather apply a so-called varying coefficients model in which the slope coefficients are unknown functions of (possibly other) covariates. The most simple case if this 'driver' and the associated covariate coincide; then you simply allow for size (or scale) effects. These considerations motivate to look for a specification incorporating more economic structure in order to induce results with more meaningful interpretation and, simultaneously flexible enough to allow incorporating regional heterogeneity, spatial dependence and nonlinearities. To this aim, this article considers as RKPF first semiparametric additive panel models, and then introduces a varying coefficient semiparametric model. Spatial dependence and spatial spillovers are accounted for with the inclusion of spatially lagged values of the knowledge inputs variables, and omitted spatial or time heterogeneity are controlled with fixed effects.

Doubtless, one can think of many different candidates to serve as so-called 'drivers' for the effect heterogeneity. Moreover, for each covariate one could choose a different one. This, however, would go beyond the scope of this paper. Instead, we concentrate - you may say 'for the sake of illustration' - on population density. An assumption underlying this choice is that population density is a good indicator for the heterogeneity of the coefficients of R&D and Human Capital, respectively, in the RKPF, and also for those of their spatially lagged counterparts. On the one hand, one may say that the relation of population density in innovation processes has not been much investigated and would deserve more attention. On the other hand, there are several indications in the literature that suggest our choice. For instance, Knudsen, Florida, Stolarick, and Gates (2008) find that the density of creative workers is a key component of knowledge spillovers and a key component of innovation. Nomaler, Frenken, and Heimeriks (2014) find a statistically significant nonlinear relation between scientific knowledge production and population density in in U.S. Metropolitan Areas. (Carlino, Chatterjee, & Hunt, 2007) show that knowledge creation measured by patent intensity is positively related to the density of employment in the highly urbanized Metropolitan Areas in US. These works recur to standard regression methods, but their findings partly indicate the existence of a role of population density on innovation activities and knowledge creation. Nonetheless, even if someone might prefer other 'drivers' for the effects' heterogeneity, the below outlined modelling approach and ideas hold equally well for any other candidate.

2.3 The Data Set

We face a panel data set covering 195 European regions from 2000 to 2012. To define regions of comparable size and governance structure, we rely on a relatively heterogeneous spatial European knowledge literature (Charlot et al., 2015; Kijek & Kijek, 2019; Bottazzi & Peri, 2003; Moreno et al., 2005b; Parent & LeSage, 2008; Paci, Marrocu, & Usai, 2014). We use NUTS 2 (Nomenclature of Territorial Units[1]) regions for Austria, Bulgaria, Croatia, Czech Republic, Finland, France, Hungary, Italy, Netherlands, Norway, Poland, Portugal, Romania, Slovakia, Spain, Switzerland and Sweden, NUTS 1 for Belgium, Germany and United Kingdom and NUTS 0 for Denmark, Estonia, Ireland, Latvia, Lithuania and Luxembourg. More details are provided in Table 2.4 in Appendix.

Table 2.1: Summary Statistics

	Variable	Mean	SD	Min	Max
Patents	K	70.574	104.58	0.02	1018.94
$R\&D$ Expenditures	$R\&D$	1.21	0.96	0.07	12.21
Human Capital	HR	34.39	9.33	11.00	63.40
Density Population	DP	324.00	797.78	3.3	7194.20

The data are downloaded from the EUROSTAT[2]. Following the above cited literature, we use as proxy for innovation, K, the number of patent applications to the European Patent Office (EPO) per million inhabitants. Similarly, the innovation inputs, $R\&D$ and HK, are respectively Research and Development expenditure measured as percentage of GDP and Human Resources in science and technology measured as share of the active population. The effects' driver, population density (DP), is measured as the number of persons per square kilometre. Summary statistics of all the variables of interest are provided in Table 2.1.

We display log of patent applications in our NUTS regions for year 2011 in Figure 2.1. One can observe a concentration of high innovation intensity in the centre of Europe (Switzerland, West of Germany and West of Austria), in Netherlands and in some Scandinavian regions, highlighting the presence of strong central-periphery distribution of innovation activity (Moreno, Paci, & Usai, 2005a). The most innovative region is North Brabant (NL) well known for its electronic activity (Philips, NXP, ASML), and the less innovative region is respectively Sud Muntenia (RO). We compute the Moran's I statistic, $I_M = 0.74$, for year 2011 to test the presence of spatial dependence in our dependent variable. This value is significant at 0.001 level rejecting the null hypothesis that innovations are randomly distributed across NUTS regions. Note that positive values of Moran's I imply positive spatial autocorrelation.

We compute the weight matrix, W, using Queen contiguity (i.e. regions are considered neighbours if they share either a side or an edge). W is then row-standardized. The finally applied Queen contiguity based neighbours are displayed in Figure 2.7 in the Appendix. Note that others type of weighting matrix are possible. We obtained relatively similar results using the inverse euclidean distance weighting matrix; see (Delgado & Robinson, 2015) for spatial weighting matrix testing.

[1] We use the version NUTS 2010.

[2] The raw data can be downloaded at `https://ec.europa.eu/eurostat/fr/home`

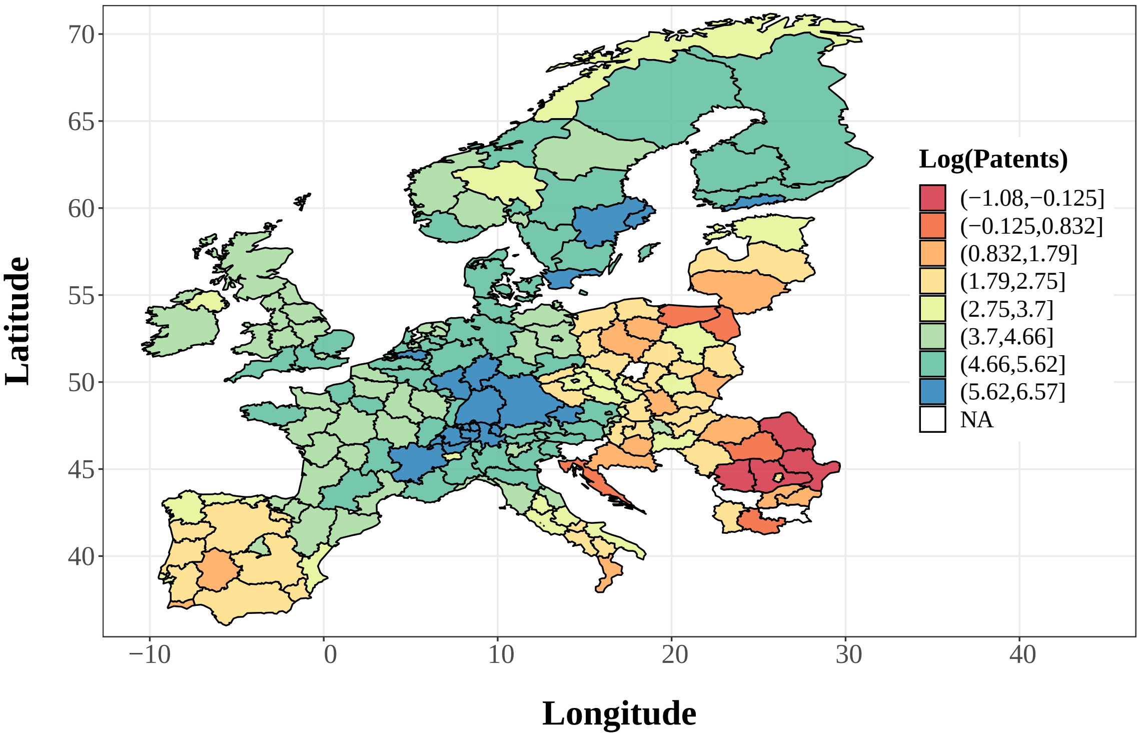


Figure 2.1: Spatial distribution of $log(Patents)$ for 2011.

2.4 A Step-wise Modelling of a Regional Knowledge Production Function with Heterogeneous Effects

In this section, we introduce step-by-step the suggested modelling approaches, always presented together the results of the estimated models plus the codes used. We concentrate here on estimation facilities provided by the free and open-source language R.

2.4.1 The classical models

We first consider the simple parametric regression model based on the traditional Griliches' KPF (Griliches, 1979) extended by the inclusion of fixed effects and of spatially lagged variables to depict knowledge spillovers. This gives, in its most general form,

$$\begin{aligned} \ln(K_{it}) &= \beta_1 \ln(R\&D_{it}) + \beta_2 \ln(HK_{it}) + \beta_3 WR\&D_{it} + \\ &+ \beta_4 WHK_{it} + \alpha_i + \delta_t + u_{it}, \quad i = 1, ..., n_t;\ t = 1, ..., T \end{aligned} \tag{2.1}$$

where $WR\&D_{it} = \sum_{j \neq i} w_{ij} \ln(R\&D_{jt})$, $WHK_{it} = \sum_{j \neq i} w_{ij} \ln(HK_{jt})$ with w_{ij} the elements of the spatial matrix W. These, like the fixed effects for region, α_i, and time, δ_t, to control for region unobserved heterogeneity and unobserved time related factors, respectively, will be included successively. More specifically, the variables $WR\&D_{it}$ and WHK_{it} are weighted averages of the i's neighbouring level of $ln(R\&D)$ and $ln(HK)$ at time t, allowing for the estimation of knowledge spillovers (indirect effects). This way is modelled the knowledge creation in region i due to the variation of the inputs of neighbours. This model can also be seen as a panel data extension of the *Spatially Lagged X-variable Model* (SLX) (Halleck Vega & Elhorst, 2015). Clearly, this is a good starting point, because among spatial econometric models it is the simplest (Gibbons & Overman, 2012). Unlike other models, no restriction on the ratio between direct and indirect effects is required (Elhorst, 2010). However, one needs to assume that the relationships between inputs and production of innovation are linear and that model's coefficients are homogeneous across regions (i.e. the effect of increasing input's level is the same across regions). These assumptions will be relaxed in the specifications further below. The results of this panel data SLX model and of differently restricted versions without spillover effects (i.e. estimating equation (2.1) without spatially lagged variables) or fixed effects are reported in Table 2.2.

Table 2.2: Estimates of the Parametric Knowledge Production Functions (2.1)

	Aspatial KPF			SLX KPF			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Intercept	-3.873^{***}			-1.394^{***}			
	(0.438)			(0.398)			
$ln(R\&D)$	1.297^{***}	0.205^{***}	0.162^{***}	0.849^{***}	0.165^{***}	0.843^{***}	0.135^{**}
	(0.043)	(0.057)	(0.058)	(0.041)	(0.060)	(0.040)	(0.059)
$ln(HK)$	2.025^{***}	1.985^{***}	1.209^{***}	1.380^{***}	1.812^{***}	1.441^{***}	1.185^{***}
	(0.124)	(0.141)	(0.199)	(0.108)	(0.159)	(0.108)	(0.199)
$WR\&D$				1.124^{***}	0.176^{**}	1.139^{***}	0.164^{*}
				(0.047)	(0.079)	(0.047)	(0.084)
WHK				-0.035	0.054	-0.017	-0.063
				(0.039)	(0.048)	(0.039)	(0.050)
Adjusted R^2	0.705	0.942	0.946	0.788	0.942	0.789	0.946
Region fixed effect	No	Yes	Yes	No	Yes	No	Yes
Time fixed effect	No	No	Yes	No	No	Yes	Yes

Note: 1473 observations *p<0.1; **p<0.05; ***p<0.01

It is well known that the inclusion of α_i and δ_t, cf. column (7), we only estimate the constant effects of the within variations in excess of the cross-region time fixed effects, those fixed effects specifications tend to underestimate the real direct effects quite importantly, see (Hauk & Wacziarg, 2009). Therefore it is recommended to countercheck those results with models including less or no fixed effects, cf. columns (4) to (6). Moreover, while for time fixed effects interpretation is often relatively clear, the regional fixed effects simply stand for any time-invariant, not-modelled heterogeneity that is correlated with 'region', they do not help for further insight into the heterogeneity like changing returns to scale, or potential effect drivers. In all models, all inputs impact either positively the production of innovation or are clearly insignificant. The most important one seems to be the endowment of human capital. Increasing the share of Human Resources in science and technology by 1%, predicts the number of patent applications to increase by up to 2%. As expected, also the $R\&D$ input is quite relevant. Increasing the share of GDP spent in Research and Development by 1% is associated with an increase of the number of patent applications by up to 1.3%. Certainly, the estimated elasticities are mitigated once spillover effects are included. In the SLX model, we distinguish direct impact and indirect impact of innovation inputs. The former corresponds to the effect of increasing

$R\&D$ and HK on the K of a given region (i.e. coefficients β_1 and β_2). The latter corresponds to the effect of increasing innovation inputs in all neighbouring regions on the K of a given region i (i.e. $\beta_3 \sum_{j \neq i} w_{ij}$ and $\beta_4 \sum_{j \neq i} w_{ij}$ which are respectively β_3 and β_4 due to row normalization of W). Joint F-tests on the significance of fixed effects exhibit their statistical significance. We observe that the inclusion of regional fixed effects decrease all coefficient values quite a bit, whereas this is much less the case for the inclusion of time fixed effects.

2.4.2 The semiparametric additive models

Certainly, one could directly switch to a fully nonparametric model. While the risk of model misspecification is minimised then, the interpretability of the estimation outcome is too. You would therefore rather try to explore the possible deviations from the benchmark specification (2.1). We first consider potentially changing returns to scale by allowing for non-linearities between innovation inputs and patent applications. This gives a nonparametric additive extension of the above SLX model(s), namely

$$\begin{aligned} \ln(K_{it}) &= f_1(\ln(R\&D_{it})) + f_2(\ln(HK_{it})) + f_3(WR\&D_{it}) + f_4(WHK_{it}) \\ &+ \alpha_i + \delta_t + u_{it}, \quad i = 1, ..., n_t; \;\; t = 1, ..., T \end{aligned} \tag{2.2}$$

where $f_1(.), f_2(.), f_3(.), f_4(.)$ are unknown smooth functions, which can take any smooth shape suggested by the data. For identification issues, these functions are typically centred to zero such that the all-over mean of $\ln(K_{it})$ is reflected in the fixed effects or an intercept if included. For the various above discussed reasons, we will estimate model (2.2) with and without the regional fixed effects α_i.[3] Note that for identification issues, you may think of asking for sufficient degrees of freedom, the more fixed effects you include, the smoother the $f_j(\cdot)$ have to be. For instance, in our case we used the mgcv package from R with penalized thin plate regression splines (Wood, 2003). Then, including the α_i requires an important reduction of basis functions, cf. the used codes shown in the Appendix. The inclusion of fixed effects in non- and semiparametric additive models has been discussed in various papers but depends strongly on the chosen smoothing method. For methods that in the moment of implementation are fully parametric, like for instance splines, this is straight forward, whereas for kernel based methods this is more involved (Profit & Sperlich, 2004; Mammen, Støve, & Tjøstheim, 2006) such that readily useable software is harder to find. The output when using splines provides also F-type statistics to test the significance of the smooth functions. Fortunately, they come to the same conclusions as the 95% confidence intervals in our figures suggest. For details on how these are calculated we again refer to Wood (2017); it is essentially a Monte-Carlos method.

The estimated smooth functions are reported in Figures 2.2 for model (2.2) with all fixed effects, and 2.3 when estimated without the α_i[4]. For the former we see an often observed phenomenon: as subjects (in our case the regions) are strongly correlated with the size of the covariates, allowing the returns to vary over size while keeping the subject specific fixed effects can produce hardly interpretable estimates. At least the general tendency of the curvatures partly correspond to our observations in Table 2.2, cf. column (7). Specifically, $\hat{f}_1$, $\hat{f}_2$ representing the relative direct impacts of $R\&D$ and HK show positive trends with $\hat{f}_2$ having a strong one, while $\hat{f}_1$ is insignificant. The estimated

[3] A residual analysis exhibited a positive correlation between residuals and time when δ_t was excluded.

[4] Note that for all provided figures, the confidence bands are pointwise.

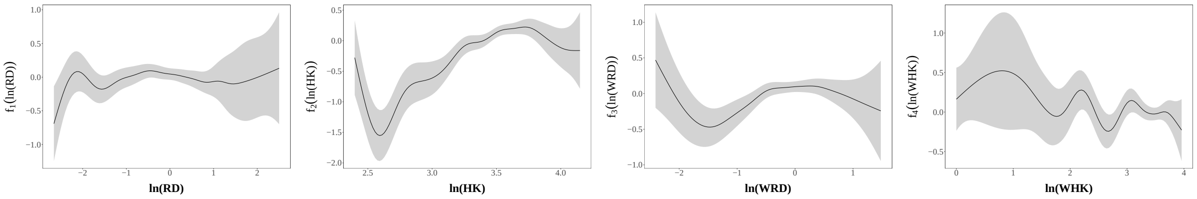


Figure 2.2: Estimated smooth functions $\hat{f}_1(\ln(R\&D))$, $\hat{f}_2(\ln(HK))$, $\hat{f}_3(WR\&D)$ and $\hat{f}_4(WHK)$ in model (2.2). Used are penalized thin plate regression splines with 15 basis functions.

indirect effect of HK, $\hat{f}_4$, has a negative trend but is insignificant. The estimated indirect effect of $R\&D$, $\hat{f}_3$, goes down, up and down again. One could also compare the coefficients shown in column (7) of Table 2.2 with

$$\frac{1}{\sum_t n_t} \sum_{t=1}^{T} \sum_{i=1}^{n_t} \frac{\partial \hat{f}_c(x_{it})}{\partial x_{it}}, \quad c = 1, 2, 3, 4 \tag{2.3}$$

with x_{ij} indicating the respective covariates. Clearly, for the common user this is only possible if that derivatives and its predictions are provided. We estimated the derivatives of smooth functions using central finite differences with the gratia package and obtain the following estimates of integrated values (2.3) respectively for f_1, f_2, f_3 and f_4: -0.03, 0.49, 0.06 and -0.09. They partly resemble to the parametric coefficient, specifically the effect of log HK.

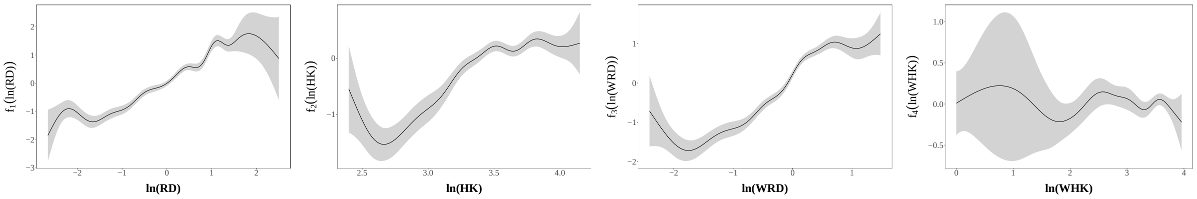


Figure 2.3: Estimated smooth function $\hat{f}_1(\ln(R\&D))$, $\hat{f}_2(\ln(HK))$, $\hat{f}_3(WR\&D)$ and $\hat{f}_4(WHK)$ in model (2.2) without α_i. Used are penalized thin plate regression splines with 15 basis functions.

As discussed, there are several reasons to countercheck the estimation results obtained with subject specific fixed effects included, recall also the critics of (Hauk & Wacziarg, 2009). If we therefore exclude the α_i, the picture changes quite a bit; see Figure 2.3. Now all counter-intuitive estimation outcomes disappear. Recall that here we are no longer estimating the within effects but explore the whole range of variation of covariates and their impact. Certainly, this is not for free as we do no longer control for potentially confounding time-invariant heterogeneity between regions. Not surprisingly, the findings resemble those of column (6) in Table 2.2, except that they show for impact functions $\hat{f}_1$ to $\hat{f}_3$ (while $\hat{f}_4$ remains insignificant) a flattening or even reversing for very small and very large values of $R\&D$, HK and WRD. Conversely, in their interquartile range, the returns are clearly higher than the parametric estimates suggest. We can again look at the integrated values (2.3), and obtain respectively for f_1, f_2, f_3 and f_4: 0.85, 0.99, 1.19 and -0.11.

2.4.3 The semiparametric varying coefficients models

Obvious disadvantages of the above approach are, among other minor issues, that the heterogeneity of effects is limited to changes returns to scale, and maybe (depends on the interest of the researcher) that for identification reasons the functions can be arbitrarily shifted and are therefore centred to zero. An interesting alternative is to return to the classical linear fixed effects model (2.1) and to think about relaxing the assumption of constant $\beta_1, \beta_2, \beta_3, \beta_4$. Well known are the options of random slopes, but these typically require independence from all covariates and u_{it}. Moreover, they show the distribution of returns but explain even less than the fixed effects α_i do. Also well studied have been time-dependent slopes by allowing all these four slopes to vary over years. More recent is the semiparametric literature on varying coefficients, see the review of Park, Mammen, Lee, and Lee (2015), and (Rodríguez Póo & Soberon, 2014) for an early approach to semiparametric varying coefficients in fixed effects panel models.

More specifically, to allow for spatial heterogeneity in the marginal effects, we extend the classical model to the Varying Coefficient Spatially Lagged X-variable model (VC-SLX),

$$\begin{aligned}\ln(K_{it}) &= f_1(DP_{it})\ln(R\&D_{it}) + f_2(DP_{it})\ln(HK_{it}) + f_3(DP_{it})WR\&D_{it} \\ &+ f_4(DP_{it})WHK_{it} + \alpha_i + \delta_t + u_{it}, \quad i = 1, ..., n_t;\ t = 1, ..., T. \qquad (2.4)\end{aligned}$$

This model can be estimated for example by using the kernel smoothed backfitting, see Roca Pardiñas, Rodríguez Álvarez, and Sperlich (2021) for the R package. The algorithm was introduced by Roca-Pardiñas and Sperlich (2010) and was based on Mammen and Nielsen (2003) and Mammen et al. (2006).

While in previous models, mainly region and time fixed effects controlled for spatial heterogeneity (by allowing for different intercepts), in the VC-SLX we would like to allow for heterogeneous coefficients. Certainly, if the heterogeneity of interest can be captured by the respective covariates, this is also true for model (2.2). Note that for the interpretation below making reference to the level of DP_{it}, its log or urbanization is essentially exchangeable.

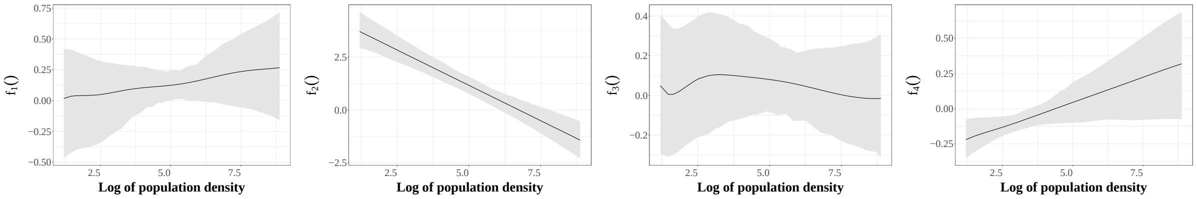


Figure 2.4: Estimated smooths coefficients of model (2.4) with all fixed effects. Used was local linear smoothed backfitting with bandwidth $h = 0.3$ for all smooth functions.

The estimated smooth coefficients are displayed in Figure 2.4 with confidence bands estimated from 200 wild bootstrap samples. The smooth coefficient associated with input $ln(R\&D)$, $\hat{f}_1(DP_{it})$, is almost always positive, clearly increasing with population density and on average > 0.15, but pretty low and flat for regions with low population density. The coefficient associated with input $ln(HK)$, $\hat{f}_2(DP_{it})$, is clearly decreasing with population density, quite high for areas with low but even negative for regions with very high population density. The latter must not be over-interpreted as at this stage we included all fixed effects, so that we are talking of within-variation effects in excess to cross-region time fixed effects. On average it is about 1.0. The smooth coefficient associated with input $WR\&D$, $\hat{f}_3(DP_{it})$, oscillates with the level of urbanization. The coefficient is always positive, on average about 0.12 and steadily on its highest levels for medium and very high population densities. Finally, the smooth coefficient associated with input WHK, $\hat{f}_4(DP_{it})$ is steadily increasing with urbanization, but on average around zero. We see that these results are not in contradiction to the fixed effects estimation in the classic linear model, recall column (7) of Table 2.2, but it gives, as expected, more insight about the heterogeneity of the covariate effects. One may also be interested in testing for constancy and linearity of the smooth coefficients following (Delgado & Arteaga-Molina, 2021) and (Mammen & Sperlich, 2022).

As in the exercises before, we countercheck these findings with estimating functions f_1 to f_4 from model (2.4) excluding α_i to exploit the covariates' full variation going beyond within effects' estimation. The results are plotted in Figure 2.5. Apart from the numerical effect reflected in different smoothness, i.e. the functions with the α_i included are less wiggly as these fixed effects filter out a lot of noise (unobserved heterogeneity), it mainly changes the scale like it happened for the classic models when comparing column (6) and (7) in Table 2.2. The only clear difference is in f_4 which becomes clearly insignificant now.

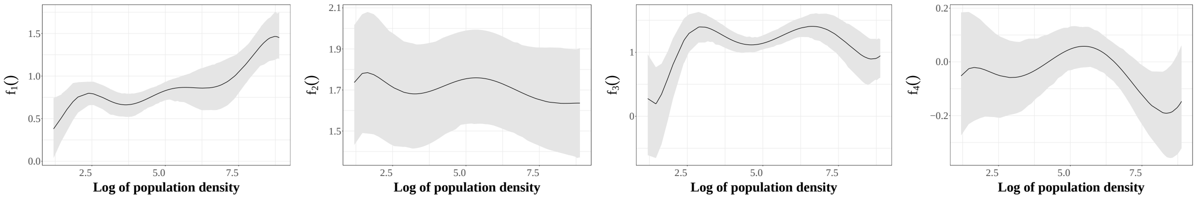


Figure 2.5: Estimated smooths coefficients of model (2.4) with year fixed effects. Used was local linear smoothed backfitting with bandwidth $h = 0.3$ for all smooth functions.

In our varying coefficient modelling approach with DP_{it} as driver, add to the standard critics against excluding or including α_i two more points: First the fact that DP_{it} hardly varies over time, and is therefore strongly (cor)related with regions. Second, one may argue that the number of patents should be transformed to a per capita indicator. Both arguments, together with the standard critics for and against subject-specific fixed effects α_i can be encountered by including nonparametrically $ln(DP)$ in our equation, namely,

$$\begin{aligned} \ln(K_{it}) &= f_1(DP_{it})\ln(R\&D_{it}) + f_2(DP_{it})\ln(HK_{it}) + f_3(DP_{it})WR\&D_{it} \\ &+ f_4(DP_{it})WHK_{it} + f_5(DP_{it}) + \delta_t + u_{it}, \quad i = 1, ..., n_t;\ t = 1, ..., T. \quad (2.5) \end{aligned}$$

Note that for reasons of optimal smoothing DP_{it} entered in log-terms in all nonparametric functions. The results are depicted in Figure 2.6. For the sake of presentation, $\hat{f}_5$ is plotted in Figure 2.8 in the appendix. As expected, it is positive throughout with a strong significantly, almost linear upwards slope.

We observe several interesting features: The estimation outcome looks very much like a compromise between the former two specifications and estimates. As expected, the slope directions remain the same, the shapes and scale are closer to those of Figure 2.5. The interpretation is still clear, although we included the $ln(DP)$ additively on the right side of the equation instead of directly looking at $ln(K/DP)$.

To summarize, the average direct effects of our covariates, defined by

$$\frac{1}{\sum_t n_t} \sum_{t=1}^{T} \sum_{i=1}^{n_t} \hat{f}_c(DP_{it}), \quad c = 1, 2, 3, 4 \qquad (2.6)$$

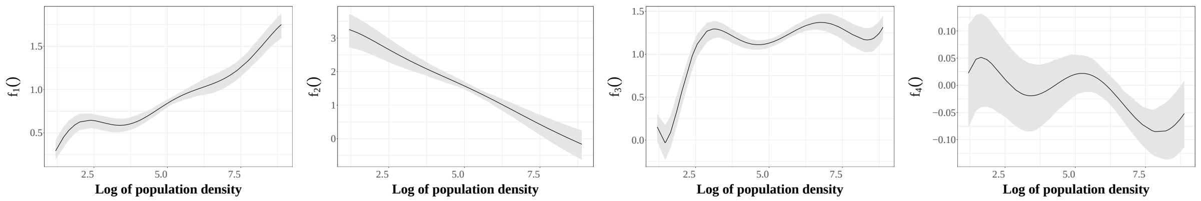


Figure 2.6: Estimated smooths coefficients of model (2.5). Used was local linear smoothed backfitting with bandwidth $h = 0.3$ for all shown smooth functions.

for the three different varying coefficient models are provided in Table 2.3. As expected, they are not equal but also not totally different from what we have in columns (6) and (7) of Table 2.2. This confirms that our specifications do not entirely change the interpretation compared to the classical approaches; however, they allow us to explore the heterogeneity of the covariates effects. Moreover, we see that sometimes heterogeneity of effects is of first order, i.e. the average direct effects - typically reported in standard empirical research - are of little political value.

Finally notice that a simple residual analysis suggests that the variance varies a bit over the NUTS regions. We therefore repeated the estimation including weights in our smoothed backfitting estimator as suggested by (Roca-Pardiñas & Sperlich, 2010). They showed that including weights inverse to the heteroscedasticity give more efficient estimates (in our case $W_{Nuts} = \hat{\sigma}^{-2}(Nuts)$). However, in our study the conclusions remain the same.

Table 2.3: Integrated values of VCM models

	VCM (2.4) with all fixed effects	VCM (2.4) with time fixed effect	VCM (2.5)
f_1	0.14	0.81	0.82
f_2	1.01	1.68	1.59
f_3	0.14	1.18	1.17
f_4	-0.04	0.00	-0.01

2.5 Final Discussion and Conclusions

In this article we are introducing and discussing various extensions of the so-called regional (or spatial) knowledge production function analysis with panel data. We have done this along European data comprising 195 European regions from 2000 to 2012. We start from some classical linear fixed effects panel model with spatial matrices to account for neighbours' impact. Then we consider mainly two different semiparametric extensions that today are readily available. Semi- and nonparametric methods have only been chosen for underlining the explorative nature of this approach. In practice, especially when facing much smaller samples, you may want to resort to these only for obtaining a better idea of model specifications suggested by the interplay of modelling and data adaptive estimation.

The different approaches are always compared to the classical ones and how we can interpret the estimates provided by those methods. We critically discuss pros and cons of each specification. Thereby we follow the spirit that each model gives a limited description of the reality, as a model always simplifies, but a good guide for model choice (apart from its statistical properties) is an appropriate compromise between interpretability and flexibility, both driven by the researcher's objective and interest. A main objective in our article is to study potential heterogeneity of the covariates' effects. Our estimates clearly indicate that their heterogeneity is indeed of first order, that is, more emphasized than the average impact itself. We believe that this is an important finding or both, a deeper understanding as well as evidence based policy.

2.6 Additional Tables and Figures

Table 2.4: Summary of selected regions

Country	NUTS	Regions
Austria	2	9
Belgium	1	3
Bulgaria	2	6
Czech Republic	2	8
Croatia	2	2
Denmark	0	1
Estonia	0	1
Finland[5]	2	4
France[6]	2	21
Germany	1	16
Hungary	2	7
Ireland	0	1
Italy[7]	2	19
Latvia	0	1
Lithuania	0	1
Luxembourg	0	1
Netherlands	2	12
Norway	2	7
Poland	2	16
Portugal[8]	2	5
Romania	2	8
Slovakia	2	4
Spain[9]	2	15
Sweden	2	8
Switzerland	2	7
United Kingdom	1	12

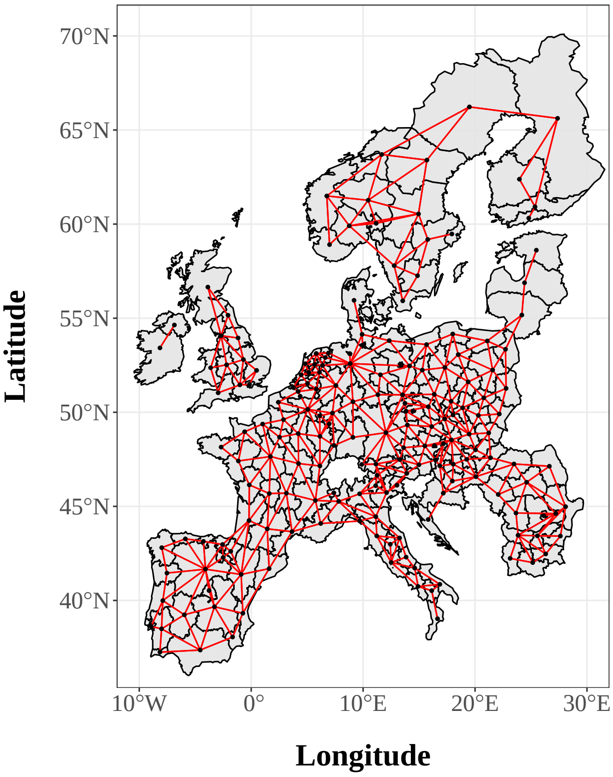


Figure 2.7: Queen contiguity based on neighbours.

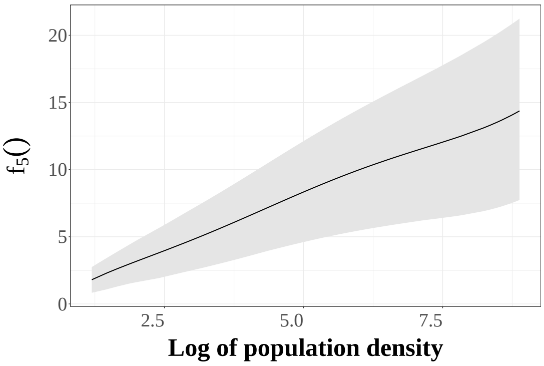


Figure 2.8: Estimated smooths function f_5 of model (2.5). Used was local linear smoothed backfitting with bandwidth $h = 0.30$.

Chapter 3

Interpretable Local Machine Learning for Huge and Distributed Data

3.1 Introduction

In his seminal talk and paper, Breiman (2001) differentiated between two cultures in the use of statistical modeling to reach conclusions from data. According to its definition, one is characterized by assuming that the data are generated by a particular stochastic data model, while the other is characterized by the use of algorithmic models in which the data mechanism is treated as unknown. He complained that the statistical community had been committed to the former one, i.e. the almost exclusive use of data models. This commitment, he argued, had 'led to irrelevant theory, questionable conclusions, and kept statisticians from working on a large range of interesting current problems'. It is evident that the second culture forms part of the more general family of machine learning methods. About 20 years later, Efron (2020) in his talk and paper, has revisited this discussion by carving out the main differences between these two cultures (see his Table 5) which, according to him, continued to develop in parallel or say, almost independently with just a few intents to close or bridge the wide gulf between these cultures. He identifies two trends for (re-)unification, one that aims to make the output of a prediction algorithm more interpretable, and one trying to achieve 'some of the advantages of prediction algorithms within a traditional framework'.

Regarding the discussion of Efron's paper, while we agree with Friedman, Hastie, and Tibshirani (2020), saying that they don't see any fundamental tension between prediction and estimation or attribution, as they all have their motivation and validity in different contexts, we agree a bit less with Yu and Barter (2020) saying that 'we are much further along the path of reunification, with the theoretical underpinnings being less critical than [...] empirical evidence in today's reality-rooted era'. Other discussants of Efron (2020) are mainly in line with his statements and conclusions. What we found a bit surprising is that in all these discussions, very little attention was given to the literature on nonparametric statistical methods with well-established theory and clear interpretation. One may either classify it as a third culture or as a main bridge pier in the middle of the above-mentioned gulf. We tend to the latter, a view that substantially motivates the analytical tool we introduce in this article. This view is intuitively supported by imagining the 'penalization' or 'smoothing parameter' of nonparametric methods as a slider to move between estimation and prediction. Let us specify the kind of nonparametric methods we are talking of.

Intuitively, high flexibility and prediction power should be achievable by high levels of local adaptiveness. We employ the plural to emphasize that the optimal level may change with location. At the same time, flexibility due to local adaptiveness typically results in estimators or predictors that are easier to understand and interpret. Consequently, we are thinking of local estimators; and to keep the presentation easy, concentrate on local smoothers. Since there exists a considerable literature about extensions of local smoothers to allow for peaks and jumps, this constraint is actually less restrictive than it seems at first; see Gijbels, Hall, and Kneip (2004), Gijbels, Lambert, and Qiu (2007), or Desmet and Gijbels (2011) for the below-considered context of local linear methods. Local adaptiveness is not only interesting for regression estimation but also for matching and causal analysis, both being specific prediction problems as highlighted in Frölich and Sperlich (2019). It is the basic principle of various estimators like for example k-nearest-neighbors (kNN), caliper and kernels, or several splines and wavelet-based estimators. Since the first two can be seen as special cases of kernel estimators, and Silverman (1984) proved the equivalence of smoothing splines with kernels, we concentrate our presentation on the latter see Schwarz and Krivobokova (2016) for the equivalence between the different spline approaches.

3.2 Local Linear Analysis in the Context of Multiple Data Sets

Consider response variables Y_i and a d-dimensional set of predictors denoted as $X_i = (X_{i1}, ..., X_{id})$, where observation pairs $\{(Y_i, X_i)\}_{i=1}^{N}$ are (typically but not necessarily) assumed to be independent and identically distributed. The easiest way to start is either to imagine that the here proposed method is an amplified local linear regression, or to think of a solution for the challenge of finding a fast algorithm that gathers in parallel from different sources (our data giants) the observations close to a given point. We rely on a fast algorithm designed to search the approximate nearest neighbors in the different large data sets (Arya, Mount, Netanyahu, Silverman, & Wu, 1998) and employ data-adaptive LASSO to select the locally optimal model(s), see Tibshirani (1996).

3.2.1 Problem Framework and Challenges

The present standard method for statistical estimation with distributed data is the Alternating Direction Method of Multipliers (ADMM) (Boyd et al., 2011) designed for distributed convex optimization, see also Parikh and Boyd (2014). This decentralized algorithm coordinates local solutions of subproblems to find the global solution. Although it converges towards the optimum, it has several drawbacks. First, convergence can be very slow, and second, optimization is often stopped earlier at intermediate solutions that are considered sufficient. Third, despite being theoretically applicable to all convex optimization problems, ADMM so far works only for simple optimization problems with little flexibility like sparse linear models (Hu, Chi, & Allen, 2016). Our method extends it significantly in several aspects while our implementation borrows some ideas of the classic ADMM. Furthermore, neither the original version of local linear estimation is designed for data giants, nor any version of LASSO is designed for local nonparametric estimation.

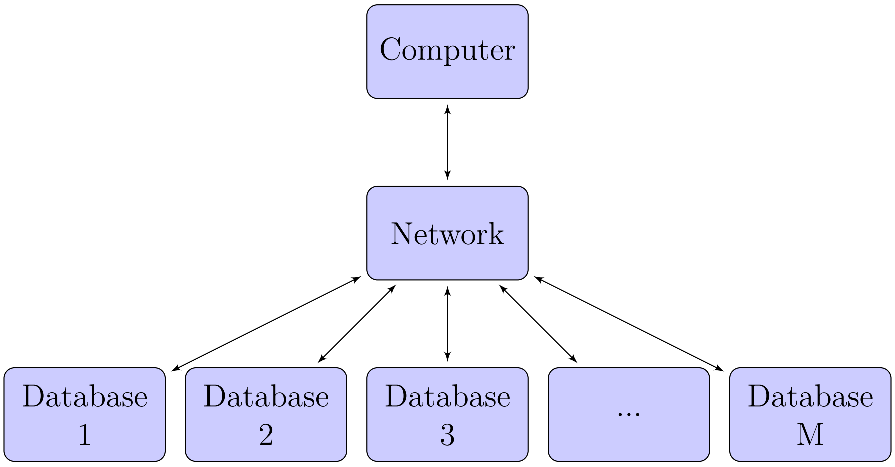


Figure 3.1: Illustration of distributed data.

In this section, we redesign and merge step-by-step all three algorithms. We do this in a way that, (a) it can be applied to distributed data, (b) without becoming a black-box, but (c) being computationally efficient, i.e. fast with only tiny approximation errors.

The distributed database system is illustrated in Figure 3.1. Data is stored in different servers that often are not physically connected to each other, though typically through a communication network. Let us focus on such database shard as it is maybe the most common system of distributed databases. Data is partitioned horizontally such that each distributed site, say *data giant*, contains a different subset of data with an overlapping set of attributes; for the sake of presentation we concentrate on these overlaps such that in each we have some observations $\{(Y_i, X_i)\}_{i=1}^{N_m}$, $N = \sum_{m=1}^{M} N_m$. The global database D is the union of data giants D_m, i.e.

$$D = D_1 \cup D_2 \cup ... \cup D_M. \tag{3.1}$$

Typically, though not necessarily, it is assumed that data are independent and identically distributed across data giants, see for example J. Zhang, Tao, and Wang (2014) for discussion.

Usually, you cannot or do not want to merge all data on one computer for legal, physical or any other reason. More generally, you may simply say that you cannot or do not want to analyze jointly all D_m, $m = 1, \dots, M$ on one processor. The so-called *divide and conquer algorithms* do then the analysis on each D_m separately and try to merge the resulting statistics to infer on D. In contrast, we are just looking for an algorithm that asks each data giant to provide the $k_m \leq k$ nearest neighbors of a point q of interest. From those, we select the k closest observations to the point of interest. As the different D_m may neither have the same distributions nor number of observations, one may criticize that we only get the k closest neighbors for sure if $k_m = k$ for all m. While this is true, though only for extremely different D_m, a resulting bias is strongly diminished by the use of kernels with falling tails (i.e. of any standard kernel except the uniform one) since this criticism concerns only observations at the boundary of the neighborhood.

For the various above described reasons, the first step is defining the local environments for which we plan to perform the local analyses to afterwards combine them via a supposed dependence structure, in our case the smoothness, recall our discussion above. After having fixed a set of points, typically a grid over the support of interest, we have to define the neighborhood for each of these points. The analysis is then conducted only with the observations being located in the respective neighborhood. There certainly exist many possibilities to define a neighborhood, but for the discussed reason we stick to kernel methods, i.e. either select all data located in a sphere centered at the point q of interest or search for the $k \in I\!N$ closest observations to this point. The latter corresponds to kNN or kernel regression with local bandwidths. In addition, it is most suitable for the implementation of a fast parallel computing algorithm.

3.2.2 Search of Nearest Neighbors From Different Sources

Thinking of large data sets, a classic kNN algorithm would be too time-consuming. Denote by N_m the number of observations in database m, d the number of predictors X_j, and $N = \sum_{m=1}^{M} n_m$ the total number of observations in the distributed system. Without loss of generality but to simplify notation, suppose that each database has about the same number of observations, so that we can suppress index m where appropriate. A naive kNN search would require at each point of interest the calculation of N distance metrics with d dimensions, and the sorting of M vectors with n entries. Even if some of these operations are not expensive, their number increases rapidly with prediction points. This suggests the use of the approximate nearest neighbors (ANN) search method of Arya et al. (1998) that significantly reduces the number of operations. Recall that the here described procedure will be applied on all data giants in parallel for $k_m \leq k$ neighbors on each, and storing their distance to the point of interest. The final k-approximate-nearest-neighbors (kANN) are those of these Mk_m neighbors with the smallest distance. For simplicity, we henceforth skip index m here.

An $(1+\epsilon)$-approximate nearest neighbor of x_0 is a point x_p whose Minkowski distance $dist(p, 0)$ to x_0 is at most by $\epsilon \geq 0$ larger than for the true nearest neighbor x_{p^*},

$$dist(p, 0) \leq (1 + \epsilon)\ dist(p^*, 0). \tag{3.2}$$

We can generalize this definition to a set of k-approximate nearest neighbors with the sequence of data points, x_{p_1}, x_{p_2}, ..., x_{p_k}, where x_{p_j} corresponds to the $(1+\epsilon)$-approximation of the jth nearest neighbors of x_0, for $1 \leq j \leq k$. The base idea is to preprocess the data into a tree structure (with cells or leaves) to report the kANN without computing a metric for all data points. Each leaf of the tree is associated with a cell containing a single data point. The ANN search starts by locating the cell in which point x_0 lies. Next, the leaf cells are enumerated in increasing order of distance to x_0, called order of priority. Leaf cells are then visited in this order, and the distance of this point in the cell to x_0 is reported. The algorithm will not visit all leaves of the tree. Inequality (3.2) defines the distance limit $l = dist(p, 0)/(1 + \epsilon)$ that stops the search, where x_p is the closest point found so far. Cells located at a greater distance from x_0 are excluded from the search. Note that increasing the relative error ϵ will speed up the algorithm because reducing l diminishes the number of cells visited. But it also increases the likelihood to miss the true nearest neighbor. Evidently, for $\epsilon = 0$, the ANN is this true nearest neighbor. The steps for the search of a single ANN are summarized in Algorithm 1.

An example of a single ANN search is provided in Figure 3.2 for an illustrative preprocessed tree. Note that the search is terminated once cell 2 is visited. Hence observation x_p is reported as the ANN of query point x_0, while the true nearest neighbor x_{p^*} is in cell 3. This highlights the trade-off implied by the choice of ϵ between speed and approximation error. The ANN search can be generalized to kANN search which aims to find the sequence $x_{p_1}, x_{p_2}, ..., x_{p_k}$. Following the same order of priority, kANN search stores the k

Algorithm 1 ANN search for query point x_0

```
Require: Preprocessed data in tree-structure; query point $x_0$; error $\epsilon$
locate $leaf_q$, leaf where $x_0$ lies
enumerate leaf cells in (increasing) order of distance from $x_0$
Start search
    visit leaf cell in the order of priority
    compute distance between single point in visited leaf cell and $x_0$
    report closest point $x_p$ found so far and $dist(p,0)$, & set $l = \frac{dist(p,0)}{(1+\epsilon)}$
    if all non-visited cells are at greater distance than $l$ from $x_0$, then stop search
    else repeat search with next leaf cell in order of priority.
return $x_p$, ANN of query point $x_0$
```

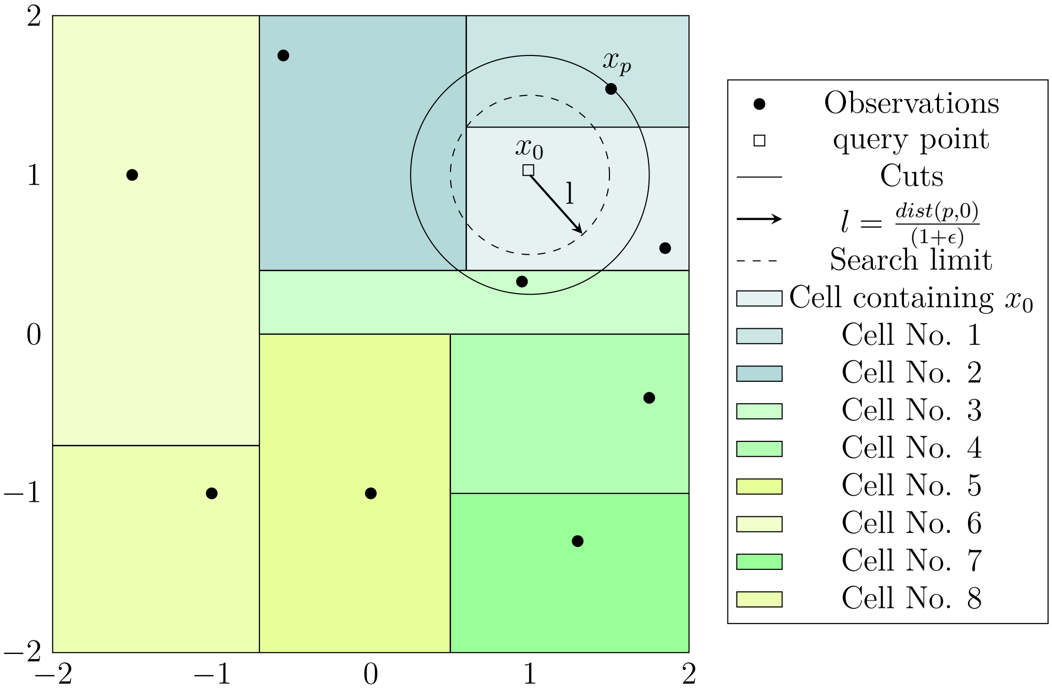


Figure 3.2: Illustration of the ANN search for a given preprocessed.

closest points found so far and computes $l_k = dist(p_k, 0)/(1+\epsilon)$. The search terminates once the distance from current cell to x_0 exceeds l_k. The ANN search allows us to find the kANN by computing distance metrics for a small amount of data. Moreover, once the data is preprocessed, the cost in time of searching kANN of additional points of interest (other x_0) is very low. This method provides a significant improvement over the naive kNN search for a moderate number of dimensions. Arya et al. (1998) recommend using it with dimensions as high as 20, but do not discourage applications with significantly higher dimensions.

There are several closely related data structures available for preprocessing the data. We chose the simplest and most widely-used one (Dasgupta & Kpotufe, 2021), the so-called k-d tree (Bentley, 1975; Friedman, Bentley, & Finkel, 1976). It is constructed by successive cuts that are placed at the median of a predictor having the highest spread in values. The space is then partitioned into 2 subspaces stored as *nodes*. This process is applied to every new node until the corresponding subspace contains a single observation. The final nodes are our leaves. The partition applied to node i is illustrated by the pre-

Preprocessed Algorithm Building a k-d tree at node i

```
Require: $S_i$, subspace at node $i$
compute $n_i$, number of observations in $S_i$
if $n_i = 1$ store $S_i$ as leaf, else
for each variable $1, ..., d$ of the subset find spread in value
select $v$, variable associated with largest spread
compute $med$, median of $v$
partition $S_i$ into two subspaces at $med$
store new subspaces as nodes
```

processed algorithm below. In the case of clustered data, the *balanced box-decomposition* (BBD) tree structure (Arya & Mount, 1995) might be more appropriate. This tree is constructed by a combination of *split* and *shrink*, where the former corresponds to k-d tree's successive cuts, and the latter partitions the space into 2 subspaces, one being inside the other. We tried different methods but in our simulations, the obtained results were independent of the chosen tree structure. We therefore chose the simplest and fastest structure to compute.

As the metric used is a Minkowski distance, we normalize our predictors (standardization is not needed) before starting the kANN search. Normalization is strongly recommended anyway for smoothing methods applied to multivariate data (Klemelä, 2009). We do this by dividing each predictor by its estimated standard deviation. Since this should not be impacted by outliers, we recommend robust estimators of the standard deviation, $\hat{\sigma}_j$. A standard choice is the median absolute deviation (MAD) popularized by Hampel (1974), namely $\hat{\sigma}_{X_j} = const \cdot median(\mid X_{ij} - X_j^c \mid)$, where X^c is an approximation of the center of the distribution, and *const* is a constant to ensure consistency. By default, the software *R* sets $const = 1.4826$, whereas X_j^c might be the median of X_j in a random subsample of D. Here we think of global outliers, referring to observations being abnormal compared to all other data in the distributed system. We must therefore compute this estimator globally and not locally or separately on each data giant. For practical reasons, we randomly select a subsample from each data giant to compute the MAD from their aggregate.

3.2.3 Local Kernel Regression With Model Selection

As above, also for the next step, it is irrelevant if we are rather thinking of a surface plus error model, or of a model-free relation between response and predictors. In either case, we are willing to relate predictor variables and response by an unknown $m(X)$ as follows:

$$Y_i = m(X_i) + u_i, \tag{3.3}$$

where u_i has $E[u_i] = 0$ and $Var[u_i] = \sigma^2(X_i)$. This presentation makes sense even if one thinks of predicting Y rather than estimating $m(X)$, because for a given x one can simply define $\hat{m}(x)$ as the predictor $\hat{y}(x)$. Our kernel method can either be seen as an estimator of a global but smooth function m or as a localization device for predicting Y with a model that is only valid at this locality around x_0. A locally parametric model suggests applying a local model selection. As the Taylor series expansion of $m(X)$ at point x_0 suggests a linear model, and the typically implemented LASSO procedures are also made for linear models, consider as objective function the following weighted least squares using only points in some neighborhood of x_0:

$$\min_{\alpha \in I\!R^{d+1}} \sum_{i=1}^{N} \left(Y_i - \alpha_0 - \sum_{j=1}^{d_1} \alpha_j (X^c_{ij} - x^c_{0j}) - \sum_{l=d_1+1}^{d} \alpha_l X^{dis}_{il} \right)^2 K\left(\frac{dm(X^c_i, x^c_0)}{max_{knn}(dm)} \right), \tag{3.4}$$

where X^c_i and X^{dis}_i indicate continuous and discrete (essentially thinking of categorical) predictors respectively. In our implementation, K is the Epanechnikov kernel and dm is the Euclidean distance computed only for the continuous predictors. Further, $max_{knn}(dm)$ denotes the maximum distance between our (approximate) k nearest neighbors and the point of interest x_0. As said, this corresponds to a local bandwidth for local linear regression for which the Epanechnikov kernel has been shown to optimize the linear minimax risk (J. Fan, 1993). The predicted response $\hat{y}_0$ is then defined as

$$\hat{y}_0 = \hat{m}(x_0) = \hat{\alpha}_0 + \sum_{l=d_1+1}^{d} \hat{\alpha}_l x^{dis}_{0l}. \tag{3.5}$$

Q. Li and Racine (2004) argue that this handling of discrete variables is not optimal and could be improved by smoothing over them too. In our framework, however, the number of discrete cells is supposed to be quite moderate compared to the sample size. As a compromise one may imagine that discrete variables with natural order get treated like continuous, and all others get decomposed into dummies with potential interactions. Note that our objective function (3.4) allows for varying coefficients for the discrete variables, i.e. to vary with x_0.

While local linear regression methods can provide good results for data sets with few continuous predictors, this is no longer the case for high-dimensional data. Already Cleveland and Devlin (1988) suggested to include variable selection in the local regression methodology. Similar to Vidaurre, Bielza, and Larrañaga (2012), we do this by adding the L_1 norm penalty to the minimization problem (3.4) to achieve a sparse solution. For a regularization parameter $\lambda > 0$, the objective function becomes then

$$\min_{\alpha \in I\!R^{d+1}} \sum_{i=1}^{N} \left(Y_i - \alpha_0 - \sum_{j=1}^{d_1} \alpha_j (X^c_{ij} - x^c_{0j}) - \sum_{l=d_1+1}^{d} \alpha_l X^{dis}_{il} \right)^2 K\left(\frac{dm(X^c_i, x^c_0)}{max(dm)} \right) + \lambda \|\alpha\|_1. \tag{3.6}$$

This regularization solves the bi-criterion problem (Boyd & Vandenberghe, 2004), where the first criterion measures the size of weighted residuals and the second measures the size of coefficients. To solve it we apply the coordinate descent algorithm (Friedman, Hastie, & Tibshirani, 2010). To see how this works, rewrite minimization (3.6) as

$$(\hat{\alpha}_0, \hat{\alpha}) = \arg\min \|W^{\frac{1}{2}}(Y - \alpha_0 \mathbf{1} - X^* \alpha)\|_2^2 + \lambda \|\alpha\|_1, \tag{3.7}$$

where W is a diagonal matrix of weights $w_i = K\left(\frac{dm(X^c_i, x^c_0)}{max(dm)}\right)$, $\mathbf{1}$ a vector of ones, and

$$X^* = \begin{pmatrix} (X^c_{11} - x^c_{01}) & \dots & (X^c_{1d_1} - x^c_{0d_1}) & X^{dis}_{1d_1+1} & \dots & X^{dis}_{1d} \\ (X^c_{21} - x^c_{01}) & \dots & (X^c_{2d_1} - x^c_{0d_1}) & X^{dis}_{2d_1+1} & \dots & X^{dis}_{2d} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ (X^c_{k1} - x^c_{01}) & \dots & (X^c_{kd_1} - x^c_{0d_1}) & X^{dis}_{kd_1+1} & \dots & X^{dis}_{kd} \end{pmatrix} \tag{3.8}$$

Denoting $\tilde{Y} = W^{\frac{1}{2}} Y$ and $\tilde{X} = W^{\frac{1}{2}} X^*$, we center all variables such that our optimization becomes

$$\begin{cases} \hat{\alpha} = \arg\min \|(I_k - \frac{1}{k}J)\tilde{Y} - (I_k - \frac{1}{k}J)\tilde{X}\alpha\|_2^2 + \lambda \|\alpha\|_1 \\ \hat{\alpha}_0 = \frac{1}{k}\mathbf{1}^t(\tilde{Y} - \tilde{X}\hat{\alpha}) \end{cases} \tag{3.9}$$

where I_k is the $k \times k$ identity matrix, and J is a $k \times k$ matrix of ones. The algorithm successively solves for $\tilde{Y}^c = (I - \frac{1}{k}J)\tilde{Y}$ the univariate minimization for each α_j $(j \neq 0)$,

$$\hat{\alpha}_j = \arg\min \|\tilde{Y}^c - \tilde{X}^c_{-j}\alpha_{-j} - \tilde{X}^c_j \alpha_j\|_2^2 + \lambda\|\alpha\|_1, \tag{3.10}$$

where $\tilde{X}^c_{-j} = (I - \frac{1}{k}J)\tilde{X}_{-j}$ correspond to the matrix $\tilde{X}^c$ with the jth column removed. Minimization (3.10) is solved effectively using the so-called *soft thresholding* function,

$$\hat{\alpha}_j = S_{\lambda^*}\left(\frac{\tilde{X}^c_j(\tilde{Y}^c - \tilde{X}^c_{-j})}{\|\tilde{X}^c_j\|_2^2}\right), \tag{3.11}$$

$$\text{with } \lambda^* = \frac{\lambda}{\|\tilde{X}^c_j\|_2^2} \quad \text{and} \qquad S_{\lambda^*}(v) = \begin{cases} 0 & \text{if } -\lambda^* \leq v \leq \lambda^* \\ v + \lambda^* & \text{if } v \leq -\lambda^* \\ v - \lambda^* & \text{if } v \geq \lambda^* \end{cases} \tag{3.12}$$

see Friedman, Hastie, Höfling, Tibshirani, et al. (2007) for details. The algorithm starts with a complete loop over all α_j and then iterates over the non-zero α's until all updates have been done. This can be done for different λ. The method we employ is called *warm start*: evaluating the α's for decreasing values of λ where the first value is $\lambda_{max} = \|X^TY\|_\infty$ setting $\hat{\alpha} = 0$. Then, minimization of (3.9) and (3.10) is done with slightly smaller values of λ such that we can use the previous $\hat{\alpha}$ as a starting value. We stop at λ_{cv}, which in turn is selected by cross-validation. This is done by launching the coordinate descent algorithm on equation (3.6) several times. At each start, a different subset of data is omitted for fitting but used to compute the squared prediction errors. The sum of these squared prediction errors are calculated for each value of lambda and finally compared to find the (optimal) λ_{cv}. The final coefficient estimates $\hat{\alpha}_{\lambda_{cv}}$ tell us which predictors are relevant at location x_0.

3.2.4 Global and/or Local Relevant Predictors

The above procedure summarized in Algorithm 2 depicts the relevant predictors locally. Imagine that in order to use the same tree structure in the kANN search for all points, you want to have the same set of predictors over the entire space. Define *global relevant predictors* as the most frequently selected variables in the space. In practice, we randomly generate several points of interest and report the selection frequencies of each predictor. The global relevant predictors are those with selection frequencies higher than a pre-specified threshold. A standard LASSO, i.e. performed on the entire data, would also have selected these global relevant predictors; but with the difference that there is no information on selection frequencies. Hence, the possibility of pre-specifying a threshold is an advantage for practitioners who want to control for the relevance of variables. Researchers investigating causes of cancer might want to set a relatively low threshold in order to minimize the risk of missing potential causes. Researchers interested in simpler but stable prediction models might want to set a high threshold to concentrate on variables with larger prediction power.

One may argue that some predictors could be relevant only in a given region of the space. The last-mentioned method will most likely define those as globally irrelevant predictors. Therefore we also provide a statistic that checks whether a predictor is only relevant in some regions, say *regionally relevant*. To describe the test, let $\mathcal{E}$ denote the set of randomly drawn points that constitute our *evaluation set* which should have

Algorithm 2 Local LASSO for distributed data

```
Require: $D$, distributed database system
Require: $k_m$, $k$ the number(s) of ANN and $T$, threshold
get MAD from subsample to normalize $X^c$    ▷ robust package (Wang et al., 2017)
draw E 'points of interest' randomly, $x_i = (x_{i1}, ..., x_{id})$, $i = 1, ..., E$
for each point of interest $x_i$ Do
    for each $D_1, ..., D_M \in D$ Do
        find $k_m$ANN of $X^c$   ▷ RANN package (Arya, Mount, Kemp, Jefferis, & Jefferis, 2019)
        compute distance $dm$ between $k_m$ANN and $x_i$
    sort $dm$ and select the $k$ smallest values
    lasso variable selection$(Y, X)$ with kANN ▷ glmnet package (Friedman et al., 2010)
    store list of relevant predictors
identify global relevant predictors $Z \subset X$ for threshold $T$
```

approximately the same distribution as the population of interest. Now think of feature X_j, i.e. dimension j of the predictor space. We denote by $\mathcal{E}_j$ the subset of all points from $\mathcal{E}$ at which dimension j was relevant along with our local LASSO. Then we check regional relevancy by comparing the distribution F_X of X which in practice is approximated by $F_{\mathcal{E}}$, with the distribution of the subset, say $F_{\mathcal{E}_j}$. If a predictor is several times selected but the two mentioned distributions are not equal, then the predictor is said to be regionally relevant. More specifically, one could test a predictor X_j for regional relevance by applying a two-samples Kolmogorov-Smirnov statistic, namely

$$KS_j = \underset{u}{sup} \mid F_{\mathcal{E}}(u) - F_{\mathcal{E}_j}(u) \mid \quad . \tag{3.13}$$

Approximate p-values are obtained as described in Marsaglia, Tsang, and Wang (2003). You may take only the global, only regional or a mixture of relevant predictors. For example, you may take the global relevant ones and add for certain regions the respective local (i.e. regionally) relevant predictors.

Large dimensions might not just be a problem inside the squared difference in (3.4), it is even more so inside the kernel function. There exists a significant amount of literature on dimension reduction to fight the so-called curse of dimensionality in nonparametrics. The literature on semiparametric modeling is abundant see for instance the recent review on generalized structured models of Roca-Pardiñas, Rodríguez-Álvarez, and Sperlich (2021), the review on specification testing or variable selection of González-Manteiga and Crujeiras (2013), the review on essential dimension reduction of Polzehl and Sperlich (2009). It is worth mentioning here that this is not in contradiction to the appetite for dimensions in some machine learning problems; it simply refers to different things: The 'curse' says that the convergence rate in estimation slows down for increasing dimensions, as for the same amount of information the complexity of the estimation problem increases. In contrast, the 'appetite' means that for the same complexity (same classification and prediction problem) the increase of dimensions may contribute additional information. The here considered LASSO (also often attributed to the machine learning tools) is usually more tailored to the former idea, motivated by the concept of potential sparsity (i.e. some predictors don't contribute information but rather add noise blurring the contribution of other predictors). In this sense, it fits well into the idea of a dimension reduction of which a kernel approach would clearly benefit (Biau & Mas, 2012).

Algorithm 3 Local linear inference for distributed data

```
Require: $D$, distributed database system and $g$, length of grid
Require: $k_m$, $k$ the number(s) of ANN and $Z$, set of global predictors
define grid with global relevant predictors, $z_i = (z_{i1}, ..., z_{ip})$, $i = 1, ..., g$
for each grid point $z_i$ Do
        for each $D_1, ..., D_M \in D$ Do
        find $k_m$ANN of $Z^c$                    ▷ RANN package (Arya et al., 2019)
        compute distance $dm$ between $k_m$ANN and $z_i$
    sort $dm$ and select the $k$ smallest values
    set $W = Epanechnikov(dm/\max(dm))$,  $\tilde{Z} = Z - z_i$ (not needed for $Z^{dis}$)
    calculate $\hat{\alpha} = (\tilde{Z}^T W \tilde{Z})^{-1} \tilde{Z}^T W Y$,  $\hat{y}_i = \hat{\alpha}_0 + \sum_{l=p_1+1}^{p} \hat{\alpha}_l \tilde{Z}_{il}^{dis}$
return graph for grid points
```

3.2.5 Final Algorithm and Remarks

Once the global relevant predictors are found, and locally some local relevant predictors added, one could update the kANN, considering only these predictors, and recomputing the now resulting Euclidean distances dm. This is in line with the literature that advises against kernel weighting with irrelevant predictors. Otherwise, an observation that is considered close to the point of interest with respect to relevant predictors could receive a too small weight if it is far from it in the irrelevant dimensions (Hall, Li, & Racine, 2007). As the sparsity assumption is more appropriate in a high dimension context, in our second round we minimize (3.4) without penalization. The main reasons are that the penalization causes a bias and that computationally we get much faster without such penalization. Update and prediction are summarized in Algorithm 3. For simplicity this is presented without the explicit adding of local relevant predictors; it is, however, obvious how to do this. Furthermore, following our suggestion to include all discrete predictors with the natural order in X^c, it is reasonable to limit the kANN procedure to X^c; recall also that only these are used for the kernel weights.

Note that we decided to rely on local λ_{cv} such that each point of interest is associated with a specific λ driven by neighboring observations. Although one may argue that a common λ_{cv} provides a middle ground between large and small penalties, we believe that neighborhoods located in different regions of the conditional distribution are not expected to share the same penalty. In other words, we want to allow the model to have locally different variation and signal-to-noise ratios. Both require to allow for locally different penalties. To the best of our knowledge, little theory has been developed so far on the impact of locally varying penalization terms; we are aware only of the paper of Krivobokova, Crainiceanu, and Kauermann (2008) who studied this phenomenon for p-splines based on linear mixed models with heteroscedastic random cluster effects.

We should finally mention the possibility of switching entirely to a LASSO procedure that is free of tuning parameters, like the TREX of Lederer and Müller (2015). What they actually do is derive and include a data-adaptive tuning parameter that has an implicit analytical expression. While intuitively this should simplify and speed up our procedure, its implementation is far from being trivial. In fact, a ready-to-use implementation of that method in either R or dynamically loadable software is not yet available, see also `github.com/muellsen/TREX`.

3.3 Simulation Results

We performed many different simulations to study the performance of kANN search, predictor selection, global vs. local λ_{cv}, computing times, and to find limits and potential problems. The largest data sets we tried had $N = 500$ million observations with only up to 10 continuous predictors, or a few million observations with up to 100 continuous predictors. Note that the number of discrete predictors plays a minor role for computational time and other problems. Note also that thanks to both, the way of managing the distributed data by parallel computing as well as our localization strategy, an increase of the absolute number N has a by far smaller impact on computational time than in standard kernel regression. More important are dimensions d and p, k_m which in our simulation we set equal to its maximum k (else you can do even much faster), size E of the evaluation sample, and grid size g. Specifically, to get an idea of the computational time for other numbers of g, you could simply take the below reported computational time of the first part and divide it by E, and the second part to divide it by g as a good approximate of the computational time per point. It is finally to be noted that the reported times are obtained with our R package which so far consists of our own, easily readable R codes combined with existing R commands of other packages. That is, on the one hand, there is still room for faster implementations by using other programming languages, on the other hand, it is still very flexible allowing for direct modification or amplification, with a maximum of adaptability and compatibility.

3.3.1 Illustration of Selection and Performance

Having said all this, it seems reasonable to limit our presentation to a simulation of a somewhat smaller scale. Specifically, we start with simulated data sets distributed over 10 data giants with each of about 1 million observations such that $N = 10{,}000{,}000$, and 20 predictors of which the last one, X_{20}, is $Bernoulli(0.5)$, and all other are independent continuous $X_j \sim N(0,1)$, $j = 1, \ldots, 19$. The response is generated as

$$Y = -X_1^2 - 2sin(\frac{\pi}{2}X_2) + X_3X_4 + \varepsilon, \qquad \varepsilon \sim N(0,1). \tag{3.14}$$

This data generating process has been chosen as it is known that, unless you have prior information about such oscillation, trigonometric functions are particularly hard to fit. A quadratic function was chosen since linear terms are nested in our local linear approach and would therefore be easily captured. Finally, we included a non-nested interaction of two globally relevant predictors. The majority of predictors is irrelevant.

We apply our method setting $k = 1{,}000$ with tolerance level $\epsilon = 1$ and $E = 1{,}000$ random evaluation points to find the relevant predictors. The selection frequencies of each predictor over the entire space are reported in Table 3.1. These results come from a single simulation but are representative as when we conducted 100 simulations, we found that the variance of selection frequencies did not exceed 0.0003. Note that we can separate the predictors in three groups: the *relevant continuous variables* with relative frequencies higher than 90%, the *irrelevant continuous variables* with relative frequencies between 25% and 35%, and the *irrelevant categorical variable* with a relative frequency below 25%. The *Kolmogorov-Smirnov* tests do not reject the null hypothesis that distributions of generated and relevant points are equal. In other words, there is no evidence that some of these predictors are (only) regionally relevant in the sense we discussed above. For any threshold T between 35% to 90% you identify X_1 to X_4 as global relevant predictors. The

Predictor	Frequencies	Two samples KS_j stat	p-value
X_1	0.955	0.017	0.999
X_2	0.950	0.024	0.941
X_3	0.924	0.006	1.000
X_4	0.924	0.005	1.000
X_5	0.330	0.070	0.187
X_6	0.328	0.048	0.625
X_7	0.331	0.038	0.866
...	...	...	...
X_{19}	0.306	0.046	0.744
X_{20}	0.212	0.014	1.000

Table 3.1: Frequency relevance of predictors. First column: available predictors, second column: frequencies of how often the variables have been chosen to be locally relevant, third column: test statistic of the two samples KS-test for local relevance, and fourth column: p-values of the KS-test.

resulting estimated conditional expectation of Y plotted on predictors X_1 and X_2 with 900 grid points is given in Figure 3.3.

We are also interested in illustrating the effect of local regularization and selection. To do so we compare the method described in Algorithm 3 with and without LASSO. There is certainly no need to show this for much smaller k. It is actually much more interesting to see if even for $k = 1{,}000$ such a local LASSO still improves on estimation.[1] Recall that an improvement is not only expected due to the reduction of local parameters to be estimated, an already well-studied phenomenon when k is not much larger than d. When $k >> d$, then it is interesting to see the effect of changing the kernel weights by dimension reduction (Vidaurre et al., 2012). This is, even more, the case in our context as d strongly affects the quality of k_m-ANN searches.

Using the data generating process described above, we predict the response value for $E = g = 1{,}000$ random points and compute the average of the Mean Squared Errors (MSE) of 50 simulations. We repeat these estimations adding successively additional irrelevant predictors. More specifically, we started with 14 predictors, namely 4 relevant and 10 irrelevant ones and increased their number up to 26. The result is plotted in Figure 3.4. As expected, the implemented local dimension reduction achieves both, reducing significantly the MSE to a number slightly below 0.005, and remaining relatively stable for increasing numbers of irrelevant predictors. In contrast, the MSE without dimension reduction steadily increases.

3.3.2 An Analysis of the Computing Time

This subsection is to study the impact of different factors and steps on computing time. It is maybe not suited for studying the speed per se, since this strongly depends on factors not directly related to procedure and algorithm (soft- and hardware, the connection of parallel processors, etc.), recall also our discussions above. All calculations are done in R with an *Acer Aspire 5* 1.8GH processor. To get an idea of the distribution of execution times inside our algorithm, we report them separately for each step of the algorithm. The easiest way is to divide the method into at least two parts: the variable selection, and

[1] We could do the same study for checking the quality of prediction.

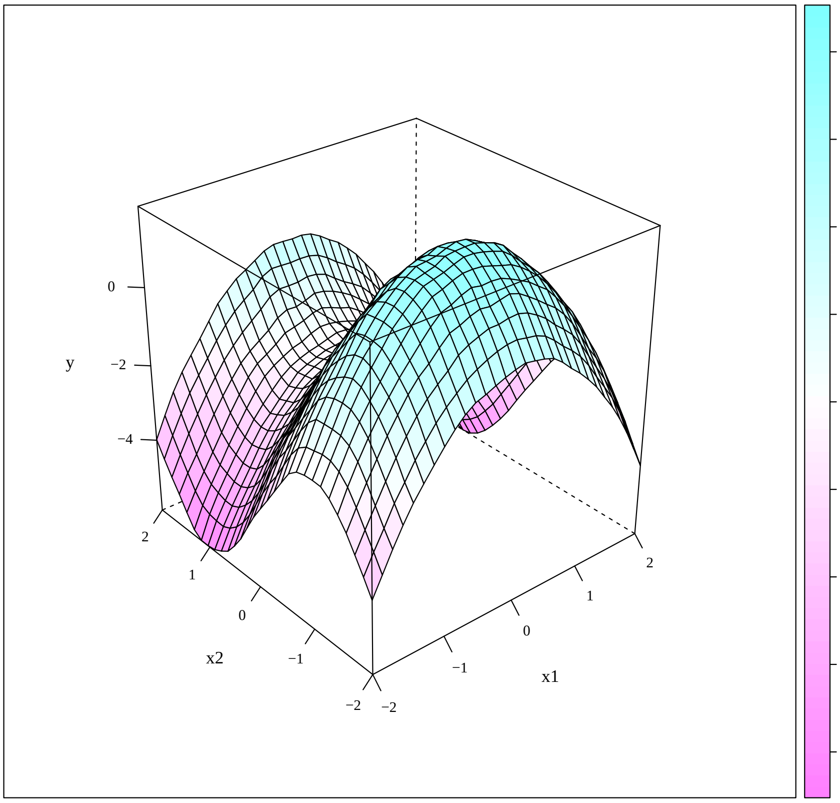


Figure 3.3: Conditional distribution of response with 900 grid points and k=1,000.

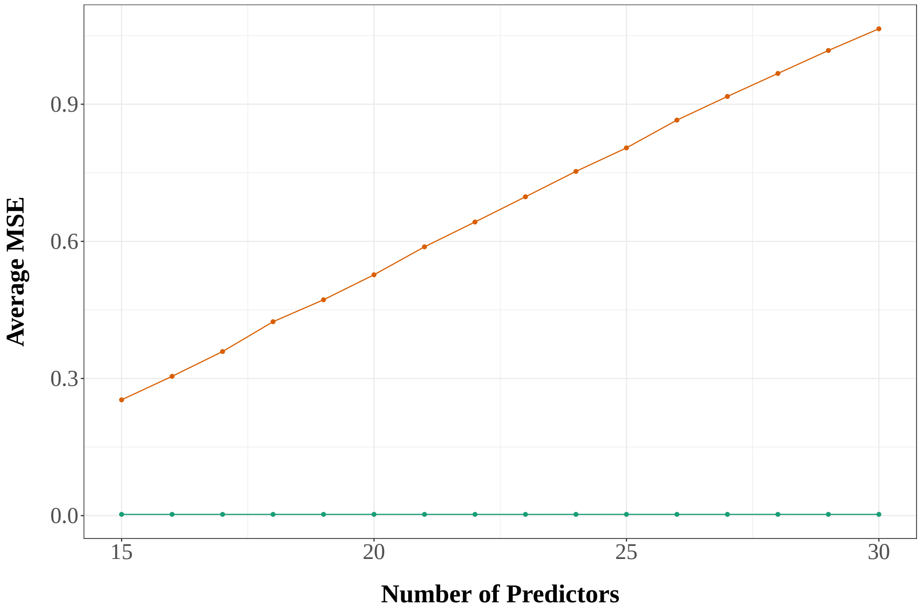


Figure 3.4: Average MSE of our procedure with (orange) and without (green) regularization for an increasing number of irrelevant predictors for $k = 1{,}000$.

the final estimation. The former is composed of the first kANN search, starting with the M parallel k_mANN searches including all predictors available, and concluding with the local LASSO which in turn includes the cross-validation for choosing the regularization parameter λ. The second part is composed of the kANN update, starting with the M

	$d = 10$			$d = 20$			$d = 30$		
	$k = 100$	$k = 500$	$k = 1{,}000$	$k = 100$	$k = 500$	$k = 1{,}000$	$k = 100$	$k = 500$	$k = 1{,}000$
First kANN search	26.88	33.55	44.23	49.59	77.12	98.51	222.6	324.32	444.93
Local Lasso	43.06	49.84	56.26	46.43	54.54	63.50	50.06	57.86	68.54
Second kANN search	20.33	24.68	31.89	22.70	28.79	38.33	23.07	30.76	36.99
Final local regression	2.88	3.47	4.63	3.93	4.98	6.36	4.89	5.85	7.57

Table 3.2: Computation times for g =900 grid points and $E = 900$ in seconds.

k_mANN searches but only for the (global or global plus some local) relevant predictors, and the final local linear regression (3.4) followed by prediction (3.5).

We reproduced the simulated example from equation (3.14) for different values of k and dimensions d, changing only the number of irrelevant predictors. The results reported in Table 3.2 are averaged over 20 simulations. Before discussing the figures plotted in the table, we need to specify their meanings and how they were calculated. First note that the procedure was run including normalization, although this was not necessary for the simulated data, to give a more realistic picture. Next, due to the independence of predictors, the program always selected the correct global relevant predictors for any reasonable threshold T. Therefore, the computing times in the second part should be theoretically independent of the original d as it only depends on p. However, as d increases, a larger part of the software memory is allocated to store the generated data which in turn increases the computation time. Further, unless the data of the different D_m are not merged on one server (for instance, in the divide and conquer context they often are, and the division is artificial), the computing times of the two kANN searches depend on aspects like M, the distribution of N_m with its different possible $k_m \leq k$ choices, the communication time between the central server and remote processors, etc. To make it independent of communication time and type of data distribution, we kept all simulated data on one computer and reported the total computing time, simply by executing the M k_mANN searches subsequently on the same processor. This certainly explains the large figures for the first kANN search. In the optimal case, you could divide them by M but have to add communication time. Recall that in our simulation we chose all N_m equal, and $k_m \equiv k$. In practice, there is some potential for saving computational time. Evidently, all numbers are also relative to the grid size g. Finally, the slowest step in a small dimension setting is the LASSO with the data-adaptive selection of the regularization parameter λ. As already mentioned, this could be accelerated by a smart implementation of TREX.

When we look at the reported figures, we first note that surprisingly, increasing the number k of neighbors does not proportionally raise the computing time of the kANN searches. As explained in Section 3.2, we are taking advantage of the fact that once the data are pre-processed, the cost of searching additional nearest neighbors is very small. Second, not surprising is that computing time is significantly decreased between the first and the second kANN search because dimension d matters a lot here. Third, the latter is much less the case for the local LASSO with an automatic λ choice; while computing time increases with k just a bit slower than it is the case for the first kANN search, it hardly increases with dimension d (compared to the kANN search).

To better understand some practical implications of Table 3.2, recall Figure 3.3. Using our R code with only one *Acer Aspire 5* 1.8GH processor, this figure was made in slightly more than 3 minutes; with $M = 10$ parallel processors, it takes about 2 minutes (depending on communication time). Have in mind that we study 10 Million observations with 20 dimensions locally adaptive (by kernels with local bandwidths) with a fully automatic

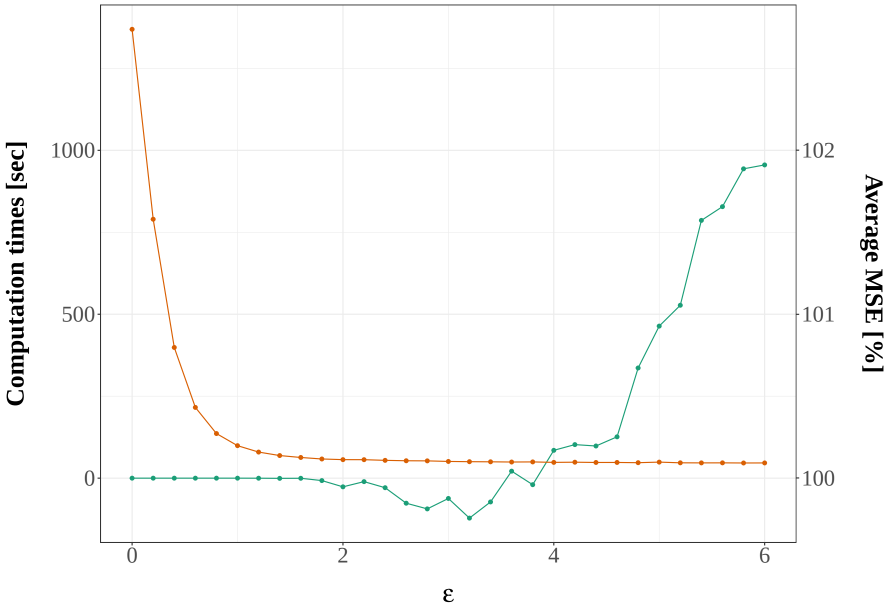


Figure 3.5: Computation times of kANN search (orange) and average MSE in percent (green).

local model selector on a relatively large grid, and without having used the full potential of time reduction e.g. by outsourcing some parts into C++ or similar.

One potential that we can check directly, apart from parallel processors, is exploring the error tolerance ϵ in the kANN search, recall Section 3.2. While there exists some theory indicating error bounds for the ϵ choice, already Arya et al. (1998) showed that these heavily overestimated the actual error levels. Larger values for ϵ imply bigger chances to select the wrong neighbors while the gains in speed can be substantial. The optimal trade-off depends on the context. In our procedure, for instance, the effect of those errors is even attenuated by the use of the Epanechnikov kernel because errors are only committed at the boundaries of the kNN neighborhoods. It is clear that the uniform or some higher-order kernels would not have such an attenuation effect.

We simulated data as above with $N = 10$ million observations, $M = 10$, $d = 20$, $k = 1{,}000$, and evaluate the final predictions at a subsample of 900 randomly drawn points by calculating their average MSE. For ϵ ranging from 0 (giving the exact kANN) to 6, Figure 3.5 shows both the average computing times in seconds of the first kANN search and the average MSE of the same 10 simulated examples for different values of ϵ. The average MSE is expressed in percentages of the average MSE when $\epsilon = 0$ is used. Figure 3.5 shows quite well why we decided for $\epsilon = 1$ as a default value: it produces essentially the same MSE as $\epsilon = 0$ but $\epsilon > 1$ hardly reduce computational time, at least not for $d = 20$. Note that in this specific simulation values of ϵ between 2 and 4 could even slightly reduce MSE. This might be because the neighborhood implied by our choice of k is smaller than the 'optimal' neighborhood. Hence, increasing the tolerance error has the effect of enlarging the neighborhood. However, for ϵ larger than 4, the error in neighbors selection is not compensated. Since the optimal choice of ϵ might depend on each specification, we decided to set a conservative tolerance level.

We finally evaluated also the scalability of our method with respect to an increasing number of observations and dimensions respectively. For the former, we repeated the same simulated example with $d = 20$ and $k = 1{,}000$, $E = 900$, $g = 900$, and $\epsilon = 1$, but increase

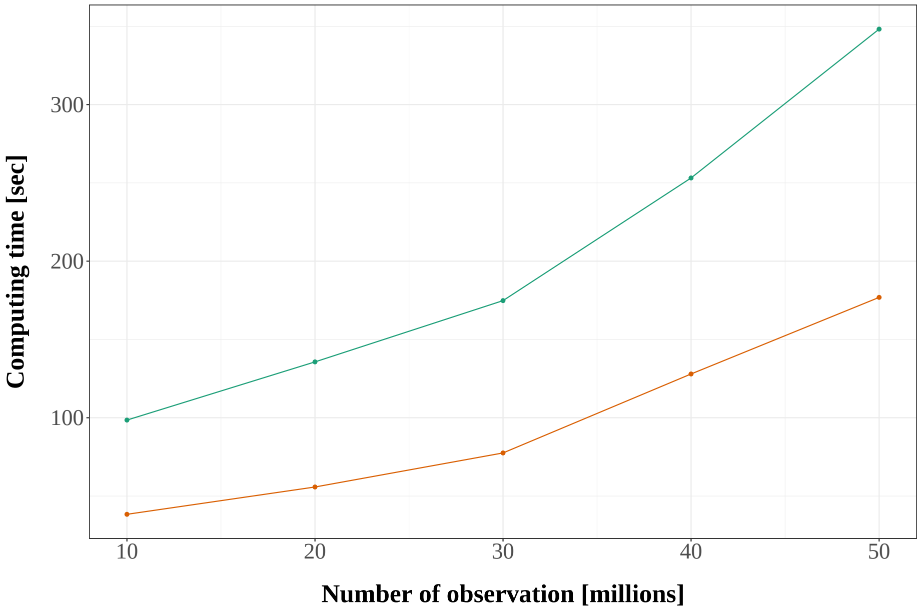


Figure 3.6: Computation times of first (orange) and second (green) kANN search.

the number of observation N of the total distributed system D from 10 to 50 millions. In Figure 3.6 we have plotted the computing times of the first and second kANN search against the increasing number of observation. As in Table 3.2 and all time figures or graphs above, the computing times for the M different k_mANN searches are added. That is, for the realistic situation in which M increases with N, as is the case with divide-and-conquer procedures, both slopes would actually become (much) flatter. For the chosen scales of minutes of calculation on millions of observations, the slope is clearly smaller than one, and far away from the exponential shape we typically observe in nonparametric inference. The latter is to be expected, since the tree structure used leads to a rate of order $O(qN \log N)$ with $q = d$ and $q = p$ in the first and second parts, respectively. This already indicates the impact of the dimension, which we study a bit further below. We also ran other simulations, not further documented, with $(k, p) = (50, 100)$ and $(k, p) = (100, 200)$, and for N much smaller. In all cases, the method performed well, accelerating significantly for smaller samples.

3.4 The Oceans' Climate Trends: An Application to in Situ Observations From a Global Array of Profiling Floats

The ocean bears the brunt of climate change as it is an important heat and carbon sink. Monitoring, modeling and predicting changes in temperature or salinity are therefore important tools to better understand climate change and its impacts on, for example, marine ecosystems (Levin & Le Bris, 2015). In this section, we briefly illustrate the potential of our proposed method in an application to a large size modern oceanography data set collected by the Argo Program (Argo, 2021).

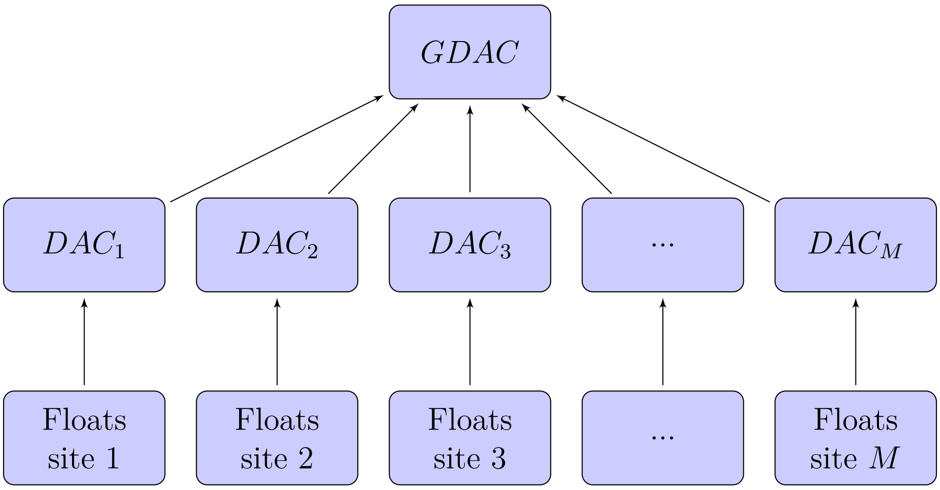


Figure 3.7: Illustration of the Argo data management system.

3.4.1 The Argo Data

Argo is a global array of more than 3,900 (active) free-floating devices ('floats') that measure the temperature and salinity of the upper 2,000 metres of the ocean.[2] The array provides over 100,000 temperature/salinity profiles and velocity measurements per year, distributed at an average spacing of 3 degrees across the world's oceans. In a 10-day cycle, the drifting floats first sink to a depth with a pressure of 2,000 dbar and then record data on temperature and salinity as they rise to the surface. At the end of each cycle, the floats send their data (referred to as 'profiles') over satellite to national *data acquisition centres* (DAC) where control tests are carried out. The data obtained is then publicly available within hours after collection via two *global data acquisition centres* (GDAC) in France and California. Figures 3.7 and 3.8 illustrate the Argo data management system and the location of national DACs. For more information on the Argo data system and its quality control procedures, and the gradual changes in the vertical resolution and spatial coverage of Argo data, see, for example, Wong et al. (2020).

Stein (2020) points out that the horizontal resolution of the array is not particularly high and space-time interpolation of Argo data is therefore of great interest. In the geoscience literature, the focus lies thus on so-called objective analysis or optimal interpolation using statistical tools like kriging (J. Li & Heap, 2008) to obtain a dense regular grid from irregularly collected data. Kuusela and Stein (2018) describe the statistical challenges associated with the use of Argo data as follows: (i) a huge volume of data;[3] (ii) the data are non-stationary in both their mean and covariance structure; and (iii) they exhibit heavy tails and other non-Gaussian features. In our application, we follow Kuusela and Stein (2018) and base our predictions (3.5) on a locally-stationary model that is only valid in the neighborhood of the point of interest. However, we do not assume the spatio-temporal mean-field to follow a Gaussian process and base our estimation on fully nonparametric techniques with locally selected predictors. Other articles in the statistical

[2] As of 17th February 2022, 3,943 active floats covered the globe. For the actual number of active floats see `https://fleetmonitoring.euro-argo.eu/dashboard?Status=Active`.

[3] As of September 2019, the data holdings in the Argo GDACs amounted to 338 gigabytes of data from 15,231 floats (Wong et al., 2020). This corresponds to around 2 million profiles, each having between 50 and 1,000 measurements.

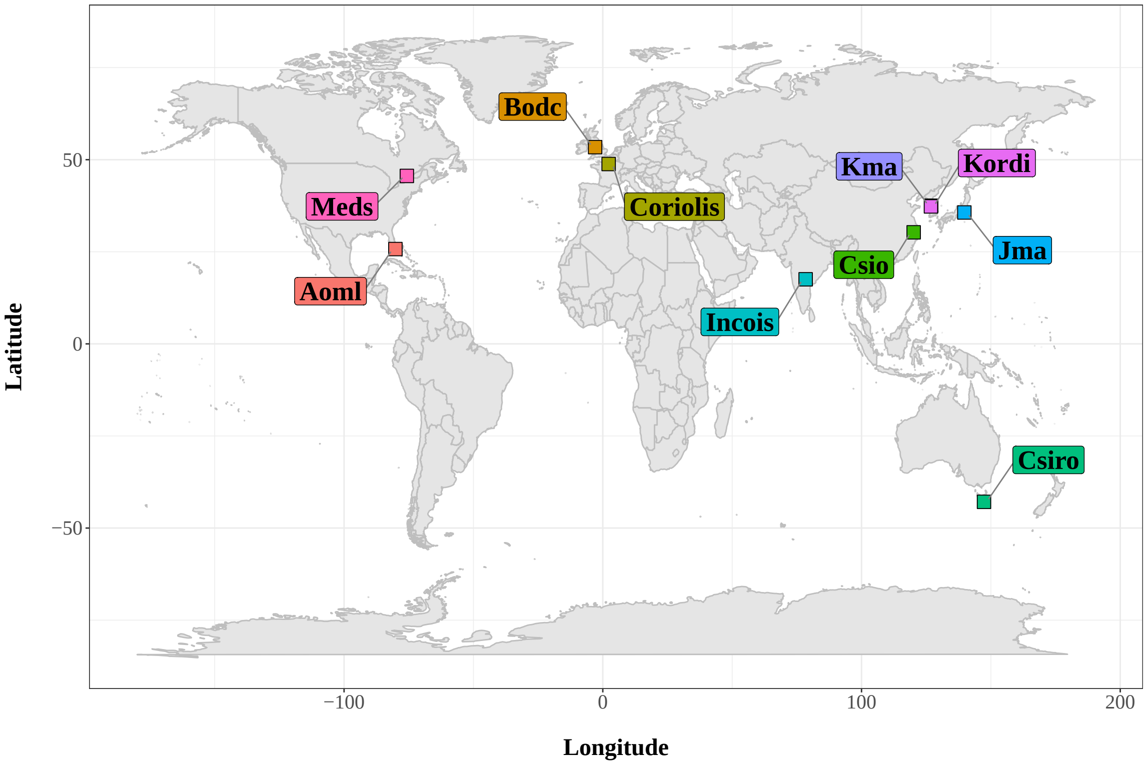


Figure 3.8: Locations of Data Acquisition Centers.

literature focus more on improvements in fitting nonstationary models to large spatial and spatial-temporal data sets (Guinness, 2021) and on the dependence of the Argo data across location, time, and pressure (Yarger, Stoev, & Hsing, 2020) using functional data analysis techniques for the problem of temperature and salinity estimation.

3.4.2 Temperature Predictions Using the Argo Data

For our illustration of temperature predictions, we follow Kuusela and Stein (2018) and Yarger et al. (2020) and use a preprocessed version of the Argo data containing more than 245 million observations for the period 2007 to 2016.[4] More specifically, we focus in our analysis on *delayed* observations corresponding to high-quality data that have been subjected to detailed scrutiny by oceanographic experts. This reduces the total sample size to about 136 million observations.

In the first step, we want to predict temperature for a grid of 9,000 locations at sea level for a specific (arbitrary) day, the 14th February 2012, based on the available predictors (salinity, pressure, latitude, longitude, as well as day and year of measurement). To further reduce complexity, we only use the part of the data that was recorded in January to March of the different years following Kuusela and Stein (2018) and divide it randomly into ten data giants of about 3 million observations each to mimic the Argo distributed system. Note that the pre-processing (the building of the k-d tree structure), the normalization, as well as the k_m-ANN searches can be done in parallel in each data giant (and in principle also remotely in each DAC). Summary statistics for the continuous variables are provided

[4] Downloaded from: https://github.com/mkuusela/ArgoMappingPaper, accessed on 07.14.2021. For the quality control criteria used for filtering out measurements due to technical issues, see the electronic supplement of Kuusela and Stein (2018).

Variable	Unit of Account	Mean	SD	Min	Max
Temperature	Degree Celsius	7.61	6.49	-1.89	33.67
Salinity	g (salt) /kg (water)	34.71	0.56	24.60	39.90
Pressure	Decibar	809.30	580.50	0.00	2,000.00
Latitude	Degree	-11.05	31.07	-72.50	79.17
Longitude	Degree	9.44	124.82	-180.00	180.00

Table 3.3: Summary Statistics of the Continuous Variables (sample size: $N = 31{,}071{,}737$)

Predictor	Frequencies	Two samples KS_j stat	p-value
Pressure	0.999	0.002	1.000
Longitude	0.899	0.009	1.000
Latitude	0.854	0.024	1.000
Day	0.806	0.013	0.999
Year	0.782	0.015	0.868
Salinity	0.280	0.161	0.000

Table 3.4: Frequency relevance of predictors

in Table 3.3.

In the first part of our procedure, $k_m = k$ is needed because the salinity and pressure values essentially separate the locations covered by the different data giants D_m. This only changes if one of these predictors is eventually identified as globally irrelevant. With $k = 1{,}000$ and $\epsilon = 1$, we obtain the selection frequencies of temperature predictors, as given in Table 3.4, based on $E = 1{,}000$ randomly drawn locations. The initial kANN search and the Local Lasso took 56.58 and 78.67 seconds, respectively, with an *Acer Aspire 5* 1.8GH processor.

Of the available predictors, only salinity is classified as a non-globally relevant predictor, having been selected as a predictor in only 28% of the cases. However, the null hypothesis of the Kolmogorov-Smirnov test is rejected (p-value=0.000), suggesting that salinity may be locally relevant. To explore this point in more detail, we show in Figure 3.9 the localization of random points where salinity is relevant (yellow) or irrelevant (blue). However, most of the points where salinity is relevant are in the Southern Ocean. This could be due to reverse causality, as higher ocean temperatures contribute to ice melt, which in turn leads to lower salinity in seawater (Pritchard et al., 2012).

After the globally relevant predictors have been selected, the temperature predictions for the grid consisting of 9,000 points can now be carried out (the second part of our procedure). The second kANN search and the final local regression took 624.22 and 37.54 seconds, respectively. Note that the significant increase in time compared to the first step is mainly due to the high memory usage for 9,000 query points. In Figure 3.10, we show our results with different values of $k \in \{100; 250; 500; 1{,}000\}$. Note that the surface temperature of the ocean is higher at low latitudes.

We also provided predictions of temperature at different depths for a fixed longitude of 30°W. Figure 3.11 shows the predicted temperature for depths ranging from 0 to 2,000 m. The very cold temperatures predicted for the latitude of -50°S correspond to the Antarctic bottom water. On the opposite pole, ocean temperatures are relatively higher due to the North Atlantic Deep Water.

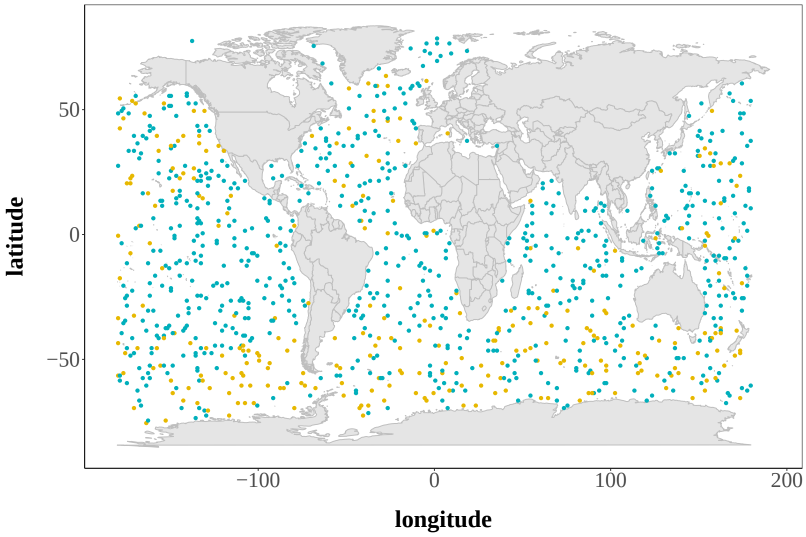
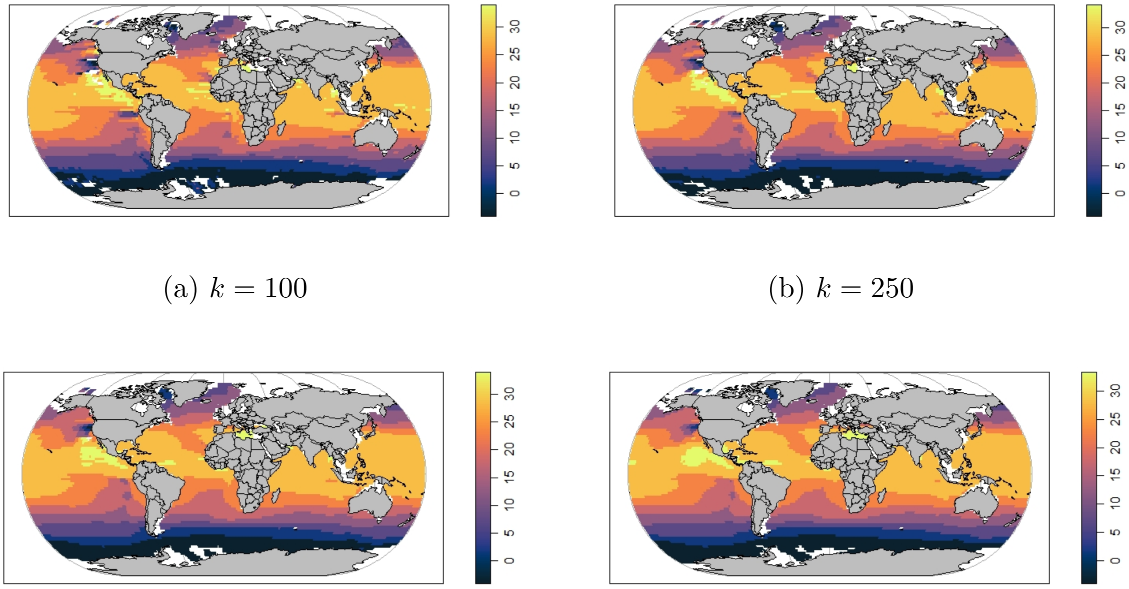


Figure 3.9: Global distribution of relevant points (yellow) and irrelevant points (blue) for salinity.



(a) $k = 100$

(b) $k = 250$

(c) $k = 500$

(d) $k = 1{,}000$

Figure 3.10: Predicted surface temperature for 14th of February 2010.

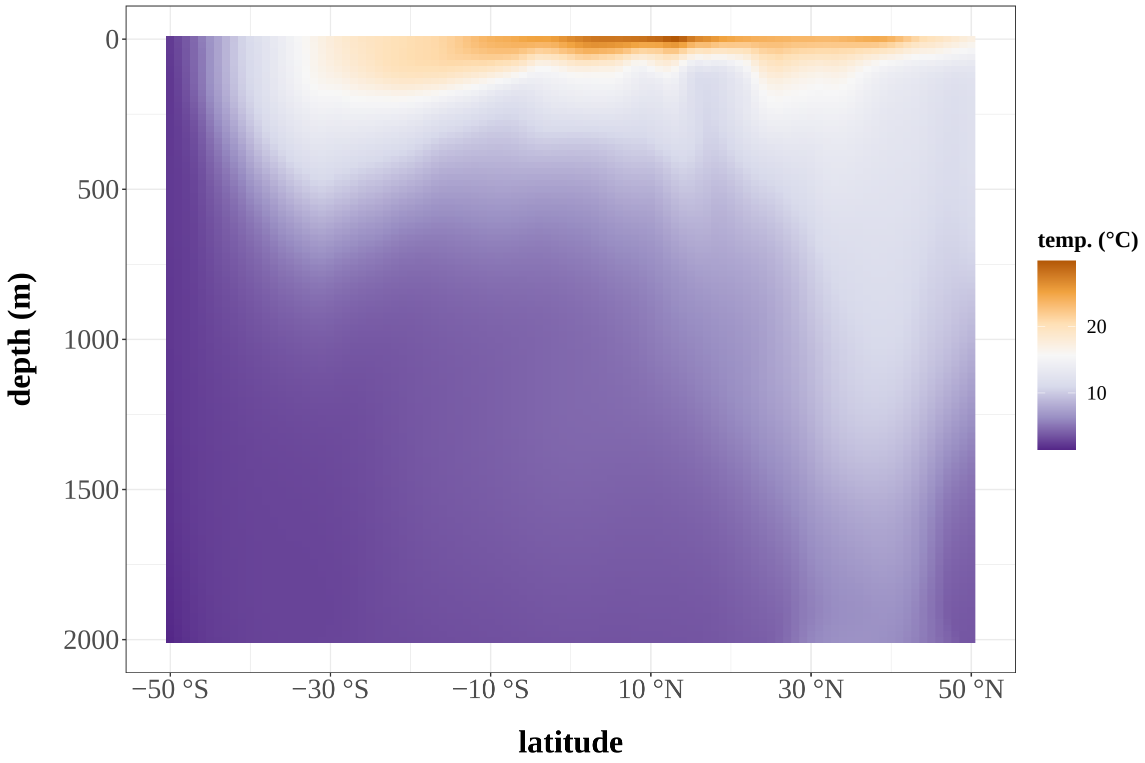


Figure 3.11: Estimated temperature for *Longitude* $= 30°$W, $k = 1{,}000$ for different depth and latitude.

In a final step, we want to investigate the warming of the ocean in the period 2007 to 2016. To do this, we predict the temperature for each of our 9,000 grid points for each day in each February from 2007 to 2016 and finally form the respective global February average temperature.[5] The results of this approach are shown in Figure 3.12, where the red line shows the temperature change at sea level and the blue line at a depth of 50 meters. The shaded areas around the curves were computed with the maximum and minimum monthly temperatures and are a measure of the variability of the predictions. Both curves show an upward-sloping trend which is more pronounced for the ocean surface. Our temperature predictions for the ocean surface are in line with the National Oceanic and Atmospheric Administration (NOAA) predicted data.[6] However, our method not only confirms the values already found for the sea surface but also shows that deeper layers follow the same trend. This provides a better understanding of the impact of climate change on the ocean.

3.5 Conclusions and Discussion

In this article, we have introduced a readily applicable procedure to perform the most flexible possible estimation, prediction and attribution related analysis on large, potentially distributed data sets. The way it is realized, our methods also contributes to the literature about 'divide-and-conquer' procedures as it might be seen as an extension of the ADMM of Boyd et al. (2011). After having revisited Breiman (2001), Efron (2020)

[5] These predictions could be done for any day of the year but we restrict our computation to February for computational reasons.

[6] `https://www.epa.gov/climate-indicators/climate-change-indicators-sea-surface-temperature`

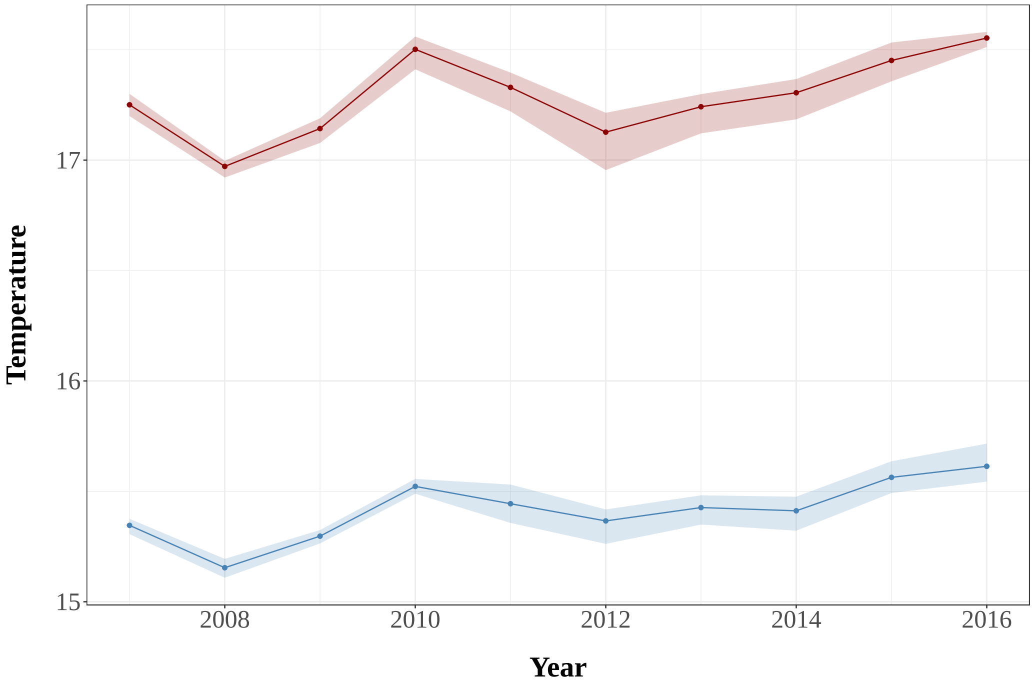


Figure 3.12: Monthly averaged temperature in degrees of February at the ocean surface (red) and 50-meter depth (blue).

and the discussions of their articles on the needs and challenges of such data analysis, we decided on the strategy of localization. Apart from the fact that classical nonparametric estimators (referring here to local smoothers) represent a pier to bridge the gulf described by Efron, it also turns banes from distributed data or divide-and-conquer problems into boons by a smart reordering of the data for parallel computing. In the spirit of strict localization, our proposition relies on the combination and adaptation of distinct algorithms that each on its own is well developed and understood: Specifically, after having organized the analysis problem as seemingly separated local problems, an efficient kANN search gathers neighboring observations for the different points of interest, a local kernel weighted LASSO (with locally adaptive penalization) infers about local attribution and sparsity, to finally return a local estimator or predictor. This is offered together with an inferential tool for distinguishing global from local relevant predictors. The specific implementations of our different steps follow computational efficiency considerations. We provide numerical examples as proof of concept. The found results confirm the use, functioning and practicability. The method is finally illustrated along with an application using the Argo project data.

What could be seen as an important distinction compared to related procedures is that albeit both characteristics, complexity and data adaptiveness, it has no feature of a black box procedure. We know perfectly what the method does to the data and how the final results are obtained. At least for the separated steps we even know the statistical properties; we admittedly cannot state analytical formulae for the final estimates for post-selection inference. Depending on the sample size, one would either apply sample splitting the way it is frequently used for many machine learning procedures, or resort to multi-fold cross-validation similar to the one applied for finding λ_{CV} above. Certainly, bootstrap procedures are feasible but computationally not very attractive in this context. As the entire procedure follows the principle of localization, we also advise against uniform inference in the sense of constructing confidence bands instead of point-wise confidence intervals in nonparametric estimation.

We conclude with a brief discussion on the distinction between estimation and prediction in the context of using our method. The distinction between traditional statistics and pure prediction algorithms is not always clear. For instance, logistic regression which belongs to traditional statistics is often present in machine learning books. In order to distinguish, Efron (2020) proposes to separate them according to their purpose. Regression approaches can be used for prediction when the aim is to predict new cases, for estimation to discern a typically smooth underlying process that generates the data, and for attribution to assign the significance of individual predictors. The distinction becomes clearer when noticing that traditional statistics are mainly used for estimation and attribution whereas pure prediction algorithms focus on prediction or classification (including clustering). To illustrate this, let's take respectively the linear regression and random forest as representatives of traditional statistics and pure prediction algorithm. The linear regression aims to investigate the true underlying data generation process through surface plus noise estimation, and uses p-values or confidence intervals for studying attribution. When linear regression is used for prediction through inter- or extrapolation of the surface, it is easily outperformed by the prediction accuracy of a random forest. Its predictive power could rely on many so-called week learners, combinations of correlated predictors, to obtain a prediction. Estimation is neglected as we cannot discern a potentially true smooth function by a black box. While one may argue that the random forest provides its own method of attribution through the so-called variance importance score, it can't determine which are the strong individual predictors when relying on many weak learners. The purpose of a random forest is then only prediction (or classification).

With a nonparametric approach like ours, you can have both, locally a clear structure with globally varying predictive power and varying contributions from, or attributions to specific predictors. Recall that most algorithms minimize a certain objective function which in our case is approximating $E[(Y - \hat{Y})^2|X]$ where $\hat{Y} = \hat{m}(X)$, see (3.3). When the information set X is allowed to be an arbitrary one and thereby arbitrarily large, this actually approximates $E[(Y - \hat{Y})^2]$; and if $m(\cdot)$ is only restricted by its smoothness in a local neighborhood, the difference between estimation and prediction shrinks to the question of what a reasonable environment is. This question, however, is an unsolved problem even within the prediction context. Therefore, in our procedure, the distinction between estimation and prediction is determined by the selection of predictors and k, where the former is guided by local LASSO with data-adaptive λ and the resulting frequencies of relevance. In contrast, we have nothing said so far about the choice of k. Recall first that we opted in favor of k instead of a bandwidth choice as we believe it complies better with the localization principle; in areas where you have more observations you are willing to relax the smoothness restriction on $m(\cdot)$. Furthermore, the problem of multivariate bandwidth choice is less accentuated when opting for elliptical kernel weighting with normalized data. Regarding the choice of k we recommend running the procedure with three to five different numbers of k (as we did in the application) to explore and better understand the local features as recommended by Chaudhuri and Marron (1999) when introducing SiZer for bandwidth selection in nonparametric kernel regression.

List of Tables

List of Figures

Bibliography

Aigner, D., Lovell, C. K., & Schmidt, P. (1977). Formulation and estimation of stochastic frontier production function models. *Journal of econometrics*, *6*(1), 21–37.

Akbari, N., Jones, D., & Treloar, R. (2020). A cross-european efficiency assessment of offshore wind farms: A dea approach. *Renewable Energy*, *151*, 1186–1195.

Aldersey-Williams, J., Broadbent, I. D., & Strachan, P. A. (2019). Better estimates of lcoe from audited accounts–a new methodology with examples from united kingdom offshore wind and ccgt. *Energy policy*, *128*, 25–35.

Anselin, L. (2010). Thirty years of spatial econometrics. *Papers in regional science*, *89*(1), 3–25.

Anselin, L., Varga, A., & Acs, Z. (1997). Local geographic spillovers between university research and high technology innovations. *Journal of urban economics*, *42*(3), 422–448.

Arbia, G. (2016). *Spatial econometrics: A rapidly evolving discipline.* Multidisciplinary Digital Publishing Institute.

Argo. (2021). *Argo float data and metadata from global data assembly centre (argo gdac)* [Dataset]. Retrieved from https://www.seanoe.org/data/00311/42182/ doi: https://doi.org/10.17882/42182

Arya, S., Mount, D., Kemp, S. E., Jefferis, G., & Jefferis, M. G. (2019). Package rann.

Arya, S., & Mount, D. M. (1995). Approximate range searching. In *Proceedings of the eleventh annual symposium on computational geometry* (pp. 172–181).

Arya, S., Mount, D. M., Netanyahu, N. S., Silverman, R., & Wu, A. Y. (1998). An optimal algorithm for approximate nearest neighbor searching fixed dimensions. *Journal of the ACM (JACM)*, *45*(6), 891–923.

Astolfi, D., Byrne, R., & Castellani, F. (2021). Estimation of the performance aging of the vestas v52 wind turbine through comparative test case analysis. *Energies*, *14*(4), 915.

Audretsch, D. B. (2003). Innovation and spatial externalities. *International Regional Science Review*, *26*(2), 167–174.

Autant-Bernard, C. (2012). Spatial econometrics of innovation: Recent contributions and research perspectives. *Spatial Economic Analysis*, *7*(4), 403–419.

Autant-Bernard, C., & LeSage, J. P. (2011). Quantifying knowledge spillovers using spatial econometric models. *Journal of regional Science*, *51*(3), 471–496.

Autant-Bernard, C., & LeSage, J. P. (2019). A heterogeneous coefficient approach to the knowledge production function. *Spatial Economic Analysis*, *14*(2), 196–218.

Bañuelos-Ruedas, F., Camacho, C. Á., & Rios-Marcuello, S. (2011). Methodologies used in the extrapolation of wind speed data at different heights and its impact in the wind energy resource assessment in a region. *Wind Farm-Technical Regulations, Potential Estimation and Siting Assessment*, 97–114.

Barros, C. P., & Antunes, O. S. (2011). Performance assessment of portuguese wind farms: Ownership and managerial efficiency. *Energy Policy*, *39*(6), 3055–3063.

Basile, R., Durbán, M., Mínguez, R., Montero, J. M., & Mur, J. (2014). Modeling regional economic dynamics: Spatial dependence, spatial heterogeneity and nonlinearities. *Journal of Economic Dynamics and Control*, *48*, 229–245.

Basile, R., & Mínguez, R. (2018). Advances in spatial econometrics: Parametric vs. semiparametric spatial autoregressive models. In *The economy as a complex spatial system* (pp. 81–106). Springer.

Benini, G., & Sperlich, S. (2021). Modeling heterogeneous treatment effects in the presence of endogeneity. *Econometric Reviews*. doi: doi.org/10.1080/07474938.2021.1927548

Benini, G., Sperlich, S., & Theler, R. (2016). Varying coefficient models revisited: an econometric view. In *Nonparametric statistics* (pp. 59–73). Springer Proceedings in Mathematics and Statistics.

Bentley, J. L. (1975, 9). Multidimensional binary search trees used for associative searching. *Communication of the ACM*, *18*, 509–517.

Biau, G., & Mas, A. (2012). Pca-kernel estimation. *Statistics & Risk Modeling*, *29*(1), 19–46.

Bosch, J., Staffell, I., & Hawkes, A. D. (2019). Global levelised cost of electricity from offshore wind. *Energy*, *189*, 116357.

Bottazzi, L., & Peri, G. (2003). Innovation and spillovers in regions: Evidence from european patent data. *European economic review*, *47*(4), 687–710.

Boyd, S., Parikh, N., Chu, E., Peleato, B., Eckstein, J., et al. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, *3*(1), 1–122.

Boyd, S., & Vandenberghe, L. (2004). *Convex optimization*. Cambridge University Press.

Breiman, L. (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science*, *16*(3), 199–231.

Buesa, M., Heijs, J., & Baumert, T. (2010). The determinants of regional innovation in europe: A combined factorial and regression knowledge production function approach. *Research policy*, *39*(6), 722–735.

Carlino, G. A., Chatterjee, S., & Hunt, R. M. (2007). Urban density and the rate of invention. *Journal of Urban Economics*, *61*(3), 389–419.

Charlot, S., Crescenzi, R., & Musolesi, A. (2015). Econometric modelling of the regional knowledge production function in europe. *Journal of Economic Geography*, *15*(6), 1227–1259.

Charnes, A., Cooper, W. W., & Rhodes, E. (1978). Measuring the efficiency of decision making units. *European journal of operational research*, *2*(6), 429–444.

Chaudhuri, P., & Marron, J. S. (1999). Sizer for exploration of structures in curves. *Journal of the American Statistical Association*, *94*(447), 807–823.

Cleveland, W. S., & Devlin, S. J. (1988). Locally weighted regression: An approach to regression analysis by local fitting. *Journal of the American Statistical Association*, *83*(403), 596–610.

Costoya, X., DeCastro, M., Carvalho, D., & Gómez-Gesteira, M. (2020). On the suitability of offshore wind energy resource in the united states of america for the 21st century. *Applied Energy*, *262*, 114537.

Crabtree, C. J., Zappalá, D., & Hogg, S. I. (2015). Wind energy: Uk experiences and offshore operational challenges. *Proceedings of the Institution of Mechanical Engineers,* *Part A: Journal of Power and Energy*, *229*(7), 727–746.

Dasgupta, S., & Kpotufe, S. (2021). Nearest neighbor classification and search. In T. Roughgarden (Ed.), *Beyond the worst-case analysis of algorithms* (pp. 403–423). Cambridge University Press. doi: 10.1017/9781108637435.024

Dee, D., Balmaseda, M., Balsamo, G., Engelen, R., Simmons, A., & Thépaut, J.-N. (2014). Toward a consistent reanalysis of the climate system. *Bulletin of the American Meteorological Society*, *95*(8), 1235–1248.

Delgado, M. A., & Arteaga-Molina, L. A. (2021). Testing constancy in varying coefficient models. *Journal of Econometrics*, *222*(1), 625–644.

Delgado, M. A., & Robinson, P. M. (2015). Non-nested testing of spatial correlation. *Journal of Econometrics*, *187*(1), 385–401.

Desmet, L., & Gijbels, I. (2011). Curve fitting under jump and peak irregularities using local linear regression. *Communications in Statistics-Theory and Methods*, *40*(22), 4001–4020.

Dismukes, D. E., & Upton Jr, G. B. (2015). Economies of scale, learning effects and offshore wind development costs. *Renewable Energy*, *83*, 61–66.

Duchon, J. (1977). Splines minimizing rotation-invariant semi-norms in sobolev spaces. In *Constructive theory of functions of several variables* (pp. 85–100). Springer.

Ederer, N. (2015). Evaluating capital and operating cost efficiency of offshore wind farms: A dea approach. *Renewable and sustainable energy reviews*, *42*, 1034–1046.

Efron, B. (2020). Prediction, estimation, and attribution. *International Statistical Review*, *88*, S28–S59.

Elhorst, J. P. (2010). Applied spatial econometrics: raising the bar. *Spatial economic analysis*, *5*(1), 9–28.

Fan, J. (1993). Local linear regression smoothers and their minimax efficiencies. *The Annals of Statistics*, *21*(1), 196–216.

Fan, Y., Li, Q., & Weersink, A. (1996). Semiparametric estimation of stochastic production frontier models. *Journal of Business & Economic Statistics*, *14*(4), 460–468.

Ferrara, G., & Vidoli, F. (2017). Semiparametric stochastic frontier models: A generalized additive model approach. *European Journal of Operational Research*, *258*(2), 761–777.

Ferreira, V., & Godinho, M. M. (2015). The determinants of innovation: a patent-and trademark-based analysis for the eu regions. In *Dynamics of knowledge intensive entrepreneurship* (pp. 393–414). Routledge.

Friedman, J., Bentley, J. L., & Finkel, R. A. (1976, 7). An algorithm for finding best matches in logarithmic expected time. *ACM Transactions on Mathematical Software*, *3*(3). Retrieved from https://www.osti.gov/biblio/1443274 doi: 10.1145/355744.355745

Friedman, J., Hastie, T., Höfling, H., Tibshirani, R., et al. (2007). Pathwise coordinate optimization. *The Annals of Applied Statistics*, *1*(2), 302–332.

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, *33*(1), 1.

Friedman, J., Hastie, T., & Tibshirani, R. (2020). Discussion of “prediction, estimation, and attribution” by Bradley Efron. *Journal of the American Statistical Association*, *115*(530), 665–666.

Frölich, M., & Sperlich, S. (2019). *Impact evaluation.* Cambridge University Press.

Giannakas, K., Tran, K. C., & Tzouvelekas, V. (2003). On the choice of functional form in stochastic frontier modeling. *Empirical Economics*, *28*(1), 75–100.

Gibbons, S., & Overman, H. G. (2012). Mostly pointless spatial econometrics? *Journal of Regional Science*, *52*(2), 172–191.

Gijbels, I., Hall, P., & Kneip, A. (2004). Interval and band estimation for curves with jumps. *Journal of Applied Probability*, *41*(A), 65–79.

Gijbels, I., Lambert, A., & Qiu, P. (2007). Jump-preserving regression and smoothing using local linear fitting: A compromise. *Annals of the Institute of Statistical Mathematics*, *59*(2), 235–272.

González-Manteiga, W., & Crujeiras, R. M. (2013). An updated review of goodness-of-fit tests for regression models. *Test*, *22*(3), 361–411.

Greene, W. (2005). Reconsidering heterogeneity in panel data estimators of the stochastic frontier model. *Journal of econometrics*, *126*(2), 269–303.

Greene, W. H. (1980). Maximum likelihood estimation of econometric frontier functions. *Journal of econometrics*, *13*(1), 27–56.

Griliches, Z. (1979). Issues in assessing the contribution of research and development to. *Bell Journal of Economics*, *10*, 92–116.

Griliches, Z. (1990). Patent statistics as economic indicators: A survey. *Journal of Economic Literature*, *28*(4), 1661–1707.

Grogg, K. (2005). Harvesting the wind: the physics of wind turbines. *Physics and Astronomy Comps Papers*, *7*.

Guinness, J. (2021). Gaussian process learning via fisher scoring of vecchia's approximation. *Statistics and Computing*, *31*. doi: 10.1007/s11222-021-09999-1

Hall, P., Li, Q., & Racine, J. S. (2007). Nonparametric estimation of regression functions in the presence of irrelevant regressors. *The Review of Economics and Statistics*, *89*(4), 784–789.

Halleck Vega, S., & Elhorst, J. P. (2015). The slx model. *Journal of Regional Science*, *55*(3), 339–363.

Hampel, F. R. (1974). The influence curve and its role in robust estimation. *Journal of the American Statistical Association*, *69*(346), 383–393.

Hastie, T. J. (2017). *Generalized additive models.* Routledge.

Hau, E. (2013). *Wind turbines: fundamentals, technologies, application, economics.* Springer Science & Business Media.

Hauk, W. R., & Wacziarg, R. (2009). A monte carlo study of growth regressions. *Journal of Economic Growth*, *14*(2), 103—147.

Heckman, J. J., et al. (1981). Heterogeneity and state dependence. *Studies in labor markets*, *31*, 91–140.

Henningsen, A. (2014). Introduction to econometric production analysis with r. *Department of Food and Resource Economics, University of Copenhagen.*

Hu, Y., Chi, E. C., & Allen, G. I. (2016). Admm algorithmic regularization paths for sparse statistical machine learning. In *Splitting methods in communication, imaging, science, and engineering* (pp. 433–459). Springer.

Hughes, G. (2012). The performance of wind farms in the united kingdom and denmark. *Renewable Energy Foundation, London.*

Iglesias, G., Castellanos, P., & Seijas, A. (2010). Measurement of productive efficiency with frontier methods: A case study for wind farms. *Energy Economics*, *32*(5), 1199–1208.

Jondrow, J., Lovell, C. K., Materov, I. S., & Schmidt, P. (1982). On the estimation of technical inefficiency in the stochastic frontier production function model. *Journal of econometrics*, *19*(2-3), 233–238.

Kang, D., & Dall'erba, S. (2016). Exploring the spatially varying innovation capacity of the us counties in the framework of griliches' knowledge production function: a mixed gwr approach. *Journal of Geographical Systems*, *18*(2), 125–157.

Kanwar, S., & Sperlich, S. (2019). Innovation, productivity and intellectual property reform in an emerging market economy: evidence from india. *Empirical Economics*, 1–18.

Kijek, A., & Kijek, T. (2019). Knowledge spillovers: An evidence from the european regions. *Journal of Open Innovation: Technology, Market, and Complexity*, *5*(3), 68.

Klemelä, J. S. (2009). *Smoothing of multivariate data: Density estimation and visualization* (Vol. 737). John Wiley & Sons.

Knudsen, B., Florida, R., Stolarick, K., & Gates, G. (2008). Density and creativity in us regions. *Annals of the association of American geographers*, *98*(2), 461–478.

Krivobokova, T., Crainiceanu, C. M., & Kauermann, G. (2008). Fast adaptive penalized splines. *Journal of Computational and Graphical Statistics*, *17*(1), 1–20.

Kuusela, M., & Stein, M. L. (2018). Locally stationary spatio-temporal interpolation of argo profiling float data. *Proceedings of the Royal Society A*, *474*(2220), 20180400.

Lederer, J., & Müller, C. (2015). Don't fall for tuning parameters: Tuning-free variable selection in high dimensions with the trex. In *Proceedings of the aaai conference on artificial intelligence* (Vol. 29).

Levin, L. A., & Le Bris, N. (2015). The deep ocean under climate change. *Science*, *350*. doi: 10.1126/science.aad0126

Li, J., & Heap, A. D. (2008). A review of spatial interpolation methods for environmental scientists. *Geoscience Australia Record*, *23*.

Li, Q., & Racine, J. (2004). Cross-validated local linear nonparametric regression. *Statistica Sinica*, *14*(2), 485–512.

Lin, B., & Luan, R. (2020). Are government subsidies effective in improving innovation efficiency? based on the research of china's wind power industry. *Science of The Total Environment*, *710*, 136339.

Lu, Z., Steinskog, D. J., Tjøstheim, D., & Yao, Q. (2009). Adaptively varying-coefficient spatiotemporal models. *Journal of the Royal Statistical Society: series B (statistical methodology)*, *71*(4), 859–880.

Mammen, E., & Nielsen, J. (2003). Generalised structured models. *Biometrika*, *90*, 551–566.

Mammen, E., & Sperlich, S. (2022). Backfitting tests in generalized structured models. *Biometrika*, *109*(1), 137–152.

Mammen, E., Støve, B., & Tjøstheim, D. (2006). Nonparametric additive models for panels of time series. *Econometric Theory*, *25*, 442–481.

Marsaglia, G., Tsang, W. W., & Wang, J. (2003). Evaluating Kolmogorov's distribution. *Journal of Statistical Software*, *8*(18), 1–4.

McMillen, D. P. (2012). Perspectives on spatial econometrics: linear smoothing with structured models. *Journal of Regional Science*, *52*(2), 192–209.

Meeusen, W., & van Den Broeck, J. (1977). Efficiency estimation from cobb-douglas production functions with composed error. *International economic review*, 435–444.

Moreno, R., Paci, R., & Usai, S. (2005a). Geographical and sectoral clusters of innovation in europe. *The Annals of Regional Science*, *39*(4), 715–739.

Moreno, R., Paci, R., & Usai, S. (2005b). Spatial spillovers and innovation activity in european regions. *Environment and planning A*, *37*(10), 1793–1812.

Myhr, A., Bjerkseter, C., Ågotnes, A., & Nygaard, T. A. (2014). Levelised cost of energy for offshore floating wind turbines in a life cycle perspective. *Renewable energy*, *66*, 714–728.

Nomaler, Ö., Frenken, K., & Heimeriks, G. (2014). On scaling of scientific knowledge production in us metropolitan areas. *PloS one*, *9*(10), e110805.

Ó hUallacháin, B., & Leslie, T. F. (2007). Rethinking the regional knowledge production function. *Journal of Economic Geography*, *7*(6), 737–752.

Olauson, J., Edström, P., & Rydén, J. (2017). Wind turbine performance decline in sweden. *Wind Energy*, *20*(12), 2049–2053.

Paci, R., Marrocu, E., & Usai, S. (2014). The complementary effects of proximity dimensions on knowledge spillovers. *Spatial Economic Analysis*, *9*(1), 9–30.

Paelinck, J. (1978). Spatial econometrics. *Economics Letters*, *1*(1), 59–63.

Parent, O. (2012). A space-time analysis of knowledge production. *Journal of geographical systems*, *14*(1), 49–73.

Parent, O., & LeSage, J. P. (2008). Using the variance structure of the conditional autoregressive spatial specification to model knowledge spillovers. *Journal of applied Econometrics*, *23*(2), 235–256.

Parikh, N., & Boyd, S. (2014). Proximal algorithms. *Foundations and Trends in Optimization*, *1*(3), 127–239.

Park, B. U., Mammen, E., Lee, Y. K., & Lee, E. R. (2015). Varying coefficient regression models: a review and new developments. *International Statistical Review*, *83*(1), 36–64.

Pinkse, J., & Slade, M. E. (2010). The future of spatial econometrics. *Journal of Regional Science*, *50*(1), 103–117.

Piribauer, P., & Wanzenböck, I. (2016). R&d networks and regional knowledge production in europe: Evidence from a space-time model. *Papers in Regional Science*.

Polzehl, J., & Sperlich, S. (2009). A note on structural adaptive dimension reduction. *Journal of Statistical Computation and Simulation*, *79*(6), 805–818.

Pritchard, H., Ligtenberg, S., Fricker, H. A., Vaughan, D. G., van den Broeke, M. R., & Padman, L. (2012). Antarctic ice-sheet loss driven by basal melting of ice shelves. *Nature*, *484*(7395), 502–505.

Proença, I., & Glórias, L. (2021). Revisiting the spatial autoregressive exponential model for counts and other nonnegative variables, with application to the knowledge production function. *Sustainability*, *13*(5), 2843.

Profit, S., & Sperlich, S. (2004). Non-uniformity of job-matching in a transition economy – a nonparametric analysis for the czech republic. *Applied Economics*, *35*(7), 695–714.

Roca Pardiñas, J., Rodríguez Álvarez, M. J., & Sperlich, S. (2021). Package wsbackfit for smooth backfitting estimation of generalized structured models. *The R Journal*, *13*(1), 330–350.

Roca-Pardiñas, J., Rodríguez-Álvarez, M. X., & Sperlich, S. (2021). Package wsbackfit for smooth backfitting estimation of generalized structured models. *The R Journal*, *13*(1), 330–350.

Roca-Pardiñas, J., & Sperlich, S. (2010). Feasible estimation in generalized structured models. *Statistics and Computing*, *20*(3), 367–379.

Rodríguez Póo, J., & Soberon, A. (2014). Direct semi-parametric estimation of fixed effects panel data varying coefficient models. *The Econometrics Journal*, *17*(1), 107–138.

Schwarz, K., & Krivobokova, T. (2016). A unified framework for spline estimators. *Biometrika*, *103*(1), 121–131.

Schweizer, J., Antonini, A., Govoni, L., Gottardi, G., Archetti, R., Supino, E., & et al. (2016). Investigating the potential and feasibility of an offshore wind farm in the northern adriatic sea. *Applied Energy*, *177*, 449–463.

Silverman, B. W. (1984). Spline smoothing: The equivalent variable kernel method. *The Annals of Statistics*, 898–916.

Soares, P. M., Lima, D. C., Cardoso, R. M., Nascimento, M. L., & Semedo, A. (2017). Western iberian offshore wind resources: more or less in a global warming climate? *Applied Energy*, *203*, 72–90.

Sperlich, S., & Theler, R. (2015). Modeling heterogeneity: A praise for varying-coefficient models in causal analysis. *Computational Statistics*, *30*, 693–718.

Staffell, I., & Green, R. (2014). How does wind farm performance decline with age? *Renewable energy*, *66*, 775–786.

Staffell, I., & Pfenninger, S. (2016). Using bias-corrected reanalysis to simulate current and future wind power output. *Energy*, *114*, 1224–1239.

Stein, M. L. (2020). Some statistical issues in climate science. *Statistical Science*, *35*, 31–41.

Stevenson, R. E. (1980). Likelihood functions for generalized stochastic frontier estimation. *Journal of econometrics*, *13*(1), 57–66.

Stone, C. J. (1986). The dimensionality reduction principle for generalized additive models. *The Annals of Statistics*, 590–606.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267–288.

Vidaurre, D., Bielza, C., & Larrañaga, P. (2012). Lazy lasso for local regression. *Computational Statistics*, *27*(3), 531–550.

Wang, J., Zamar, R., Marazzi, A., Yohai, V., Salibian-Barrera, M., Maronna, R., . . . others (2017). Package robust.

Weinstein, M. A. (1964). Query 2: The sum of values from a normal and a truncated normal distribution. *Technometrics*, *6*(1), 104–105. Retrieved 2022-06-03, from http://www.jstor.org/stable/1266751

Wen, Y., Kamranzad, B., & Lin, P. (2021). Assessment of long-term offshore wind energy potential in the south and southeast coasts of china based on a 55-year dataset. *Energy*, *224*, 120225.

Wong, A. P. S., Wijffels, S. E., Riser, S. C., Pouliquen, S., Hosoda, S., Roemmich, D., . . . Park, H.-M. (2020). Argo data 1999-2019: Two million temperature-salinity profiles and subsurface velocity observations from a global array of profiling floats. *Frontiers in Marine Science*, *7*. Retrieved from https://www.frontiersin.org/article/10.3389/fmars.2020.00700 doi: 10.3389/fmars.2020.00700

Wood, S. (2003). Thin plate regression splines. *Journal of the Royal Statistical Society, B*, *65*(1), 95–114.

Wood, S. (2017). *Generalized additive models – an introduction with r*. Chapman and Hall/CRC.

Wood, S., & Wood, M. S. (2015). Package 'mgcv'. *R package version*, *1*, 29.

Yarger, D., Stoev, S., & Hsing, T. (2020). A functional-data approach to the argo data. *ArXiv Preprint arXiv:2006.05020*.

Yu, B., & Barter, R. (2020). The data science process: One culture. *Journal of the American Statistical Association*, *115*(530), 672–674.

Zhang, F., Wang, Y., & Liu, W. (2020). Science and technology resource allocation, spatial association, and regional innovation. *Sustainability*, *12*(2), 694.

Zhang, J., Tao, X., & Wang, H. (2014). Outlier detection from large distributed databases. *World Wide Web*, *17*(4), 539–568.