Ph.D. Research Proposal
Geographical Applications of modern statistical learning algorithms

Ron Sarafian

Ben Gurion University

The Faculty of Engineering Sciences

Department of Industrial Engineering and Management

Advisor: Dr. Johnathan Rosenblatt

Advisor: Prof. Itali Kloog

Advisor: Prof. Israel Parmet

Contents

1 Abstract in Hebrew

2 Abstract

Estimating air pollution concentrations is valuable for environmental exposure assessment and epidemiological studies. The use of satellite-based data models to estimate particulate matter (PM) has increased substantially over the past few years. These models employ statistical learning algorithms to analyze spatially and temporally resolved datasets to provide an assessment of air pollution levels in locations where no measurement is performed. Assessments are then used as covariates in epidemiological studies; hence, their reliability is of high importance.

However, some complex issues arise when analyzed data is characterized by spatial and temporal structure, many of which are neither clearly defined nor completely resolved. Spatio-temporal dependency is a challenge for classical statistical methods, especially those used for prediction and model performance assessment. Therefore, proper statistical techniques that reduce prediction's errors are necessary in order to avoid biased epidemiological results that may lead to erroneous conclusions. scale

Our research focuses on both theoretical and practical aspects of modern statistical learning algorithms implemented in environmental and geographical applications. It is carried out as a research collaboration between the Department of Industrial Engineering and Management and the Department of Geography and Environmental Development under the supervision of Dr. Johnathan Rosenblatt and Prof. Itai kloog.

We study the statistical issues regarding the estimation and model assessment of air pollutants concentrations in the presence of spatio-temporal autocorrelation in the data. In particular, the research aims at: (i) Propose an extension to classical methods that estimate prediction error (such as Cross Validation) for exposure assessment models with spatio-temporal structured data. (ii) Improve models prediction performances while reducing its complexity using linear models. (iii) Cope with computational challenges associated with these models, taking advantage of some recent significant achievements in the fields of memory efficiency and parallel computing.

We hope that this research will greatly contribute to the fields of environmental exposure and geo-statistics; not only from theoretical perspective, but also in that it would harness scientific knowledge to produce applications that would improve people's life.

3 Introduction

How does air quality affect people health? A quantitative answer for this question is of interest for policy makers, especially regarding to the allocation of pollutants. To provide a general answer, epidemiologists usually employ observational studies, in which the relationship between air pollution expusure level and health indices is being investigated. In many cases the

experimental units are spatio-temporal entities, each with its own measurement of pollution concentration. The fine *Particular Matter* (PM) abundance is one of the regularly monitored air pollutants. However, PM mass concentrations are measured at ground monitoring stations, therefore highly limited in terms of spatial coverage. In order to provid PM assesments in geographical areas where no mesurments are available some statistical methodologies are needed.

PM levels show high spatio-temporal variation (Pelletier, Santer, and Vidot 2007), therfore, geostatistical interpolation methods in which spatial PM levels are modeled as opposed to some smoothness minimizer polynomial spline may introduces PM exposure error (Zeger et al. 2000). Over the years, methods which utilize additional data to minimize the exposure error have been developed. The *Land-use regression* (LUR) approach takes advantage of traffic, topography, and other geographic variables and train models to predict pollution levels at any location (Briggs et al. (2000); Yanosky et al. (2008)). The LUR methods provided more accurate exposure assessments at unmonitored locations. However they were significantly limited mostly since that included covariates are generally not time varying, and therefore were better suited for long-term exposure assessments. Along with the improvement in the epidemiologists' analysis tools the demand for spatio-temporally resolved datasets of PM concentrations was growing and new approaches became inescapable.

The increase availability of satellite data has enabled environmental scientists to harness remote sensing technologies to PM concentration assesments. The satellite-based *Aerosol Optical Depth* (AOD) retrievals were found to be associated with ground PM measurments in different geographical areas. It is the measure of particles in air (e.g., haze, smoke, dust) distributed within a column of air from the earth's surface to the top of the atmosphere, so it serves as a good proxy for PM. Due to its large temporal and spatial coverage, the AOD allows to predict the PM levels in times and spaces where surface PM measurements are not available.

Indeed, the use of AOD as a covariate in studies of PM concentration estimation has increased substantially over the past few years. In recent studies, ground measured PM data were integrated with rerpeated AOD measurments along with other spatio-temporal covariates in order to build a model that can be used to assess PM exposure in locations where no measurement is performed. The clustered structure of the data requires a model that is able to account for several sources of variability (mostly, space and time), and a common practice is the *Mixed models* statistical framework (Kloog et al. (2011); A. A. Chudnovsky et al. (2014) Kloog et al. (2015); M. Lee et al. (2016)). The AOD based models have shown substantially well performances in predicting exposure levels, and their great advantage is in providing exposure assessments within short intervals.

The generated PM predictions allow to examine the effect of air pollution on health and disease conditions, not only near monitoring stations, but also in remote, usually less populated areas. Studies that have used generated PM predictions which were obtained using the abovementioned procedure found statistically significant correlation between PM and varaity of adverse health effects such as birth outcomes (Kloog et al. 2012); natural-cause mortality (Y. Wang et al. 2016); acute myocardial infarction Madrigano et al. (2013)] and more.

The rest of this proposal is organized as follow: Section 4.1 introduces some of the challenges in satellite-based PM assessments models. Section 4.2 presents the research objectives that are associated with these challenges. Section 5.1 provides a detailed review of the validation methods appropriate for the data with dependent structure that can be developed and fit. Section 5.2 discusses the model statistical framework and presents a generalized modification. Section 5.4 shows how some of the achievements in the fields of memory efficiency and parallel computing can be implemented in our proposed framework and improve computational performance.

4 Study Approach

4.1 Challenges

The mission of PM exposure levels assesment is actually not well defined without additional settings. Put differently, the study meta-purpose must be declared. This purpose is of great importance since it derives the appropriate performance evaluation methodology. An old proverb says that *"The proof is in the pudding"*, therefore, a significant decision one should make is which pudding to choose. Should the assessed PM values minimize the mean squared prediction error? Perhaps a more conservative approach is appropriate and the maximum error should be minimized? Obviously, these choices should be considered in light of the epidemiological study goals. For instance, when estimating the effect of air pollution on health condition, should the emphasis of PM assesment be placed on being as accurate as possible in populated areas, rural areas, or evenly across space?

Moreover, even after the study purpose was determined, executing model performance evaluation is not a trivial task. The spatial and temporal structure of the analyzed data raises some complex statistical issues, many of which are neither clearly defined nor completely resolved. In particular, when spatio-temporal dependency exist, assessing how the results of the model will generalize to an independent data set becomes a challenging problem that is less explored. *Cross-validation* (CV) is a widespread strategy that is used for model's predictive performance estimation. The main idea behind CV is to avoid overfitting by splitting the data several times into training and validation samples which are independent. However, when data are spatio-temporaly dependent, classical CV analyses break down and must be modified.

Anoter challenge which arises from the structure of the data is related to statistical model that is being used. Providing PM assesments in geographic locations and at different time points requires the learning from complex dynamic spatial datasets. These datasets have a natural hierarchical or multi-level structure. On one hand, it offers researchers extended modeling possibilities, thus to increase model performance. On the other hand, the analysis of these datasets requires a proper modeling of the observations dependence structure. There are several modeling approaches, each with its own pros and cons. Obviously, any selected approach has its own influence on the epidemiological resuls.

When databases are also very large, the computational complexity of the model is additional consideration should be taken into account. Different models usually have different estimation procedures with its own properties of computational complexity. For instance, the mixed models are usually estimated using *Maximum Likelihood* (ML) estimation method, therefore are based on numerical optimization of nonlinear functions with no closed-form solution. If on the other hand, a linear model is chosen, there is a closed-form analytical solution, which generally is easier to compute. Moreover, computational hurdle might be reduced by efficient representation of the data, as well as implementation of parallel computing mechanism.

4.2   Research Objectives

The research objectives follows the challenges that were described in section 4.1. In principle, the order of tackling the objectives is as it appears in the following sections, however, there are plenty of overlaps subjects, so that progress is expected in several channels in parallel. The objectives are:

- Investigate the procedure of PM prediction model performance estimation with CV in conditions of spatio-temporal dependency and in light of the epidemiological goals.

- Propose a generalization to the state-of-the-art models that are used to predict PM levels in order to improve their performances. We will examine the *Genralized Least Squares* (GLS) model and suggest several methods for estimating its required covariance matrix while modeling the spatio-temporal dependencies. A regularization of the GLS model would also be investigated.

- Present an improvement in model's computation time and memory use, taking advantage of the fact that linear models consume less compututional resources. We would also employ some of the recent achievements from the fields of sparse representation, parallel computing and memory efficiency to reduce computational hurdle.

5   Research Framework

5.1   Performance Estimation

There are many validation techniques for evaluating the predictive performance of a statistical algorithm. Probably the simplest and most widely used is *Cross-validation* (CV). The main idea behind CV is as follow: First, split the data into a training set and a validation set. It is crucial that these sets are completely independent, otherwise CV is likely to be overfitting. Second, use the training set to train a statistical algorithm. Third, evaluate the predictive performance of the trained algorithm on the validation set.

In Spatio-temporal autocorrelated settings, it is fundamental to find a strategy that split the data into two samples that would be independent, yet would not unwittingly induce extrapolations by restricting unnecessary range between training and validation samples.

Forthermore, an obvious question is how to evaluate the predictive performance. In a regression setting for example, the performance of a model is measured by the discrepancy between the real and the predicted values, often in a form of the *sum of squers*. Following the framework of Arlot, Celisse, and others (2010), the quality of the predicted values as an approximation to the real values is quantified by its loss $\mathcal{L}$. We will later discuss the exact settings of the loss function $\mathcal{L}$, but before that we would like to emphasize the importance of choosing it.

Supervised learning tasks such as regression can be formulated as the minimization of a loss function over a training set. Thus, the choice of the loss function should reflect the purpose of the study. For instance, if the epidemiological objective is to estimate the effect of air pollution on health condition, then PM prediction by its own is not the ultimate goal. In this case, choosing the loss function is not trivial. It is necessary to find the loss function which through its minimization the estimated effect of air pollution on health is the closest to the truth.

It seems that when one is willing to perform a CV procedure in this settings, he is facing two different missions: (i) choosing the right loss function so that it meets the objectives of the study, and (ii) choosing the folding scheme so that CV overfitting due to spatio-temporal dependency is prevented. We will address each mission separately.

5.1.1 The Loss Function

In order to illustrate the meaning of choosing the loss function and its influence on epidemiological results, let us discuss an example of an extreme situation: Consider the case where 100 PM monitoring stations are scattered in some geographical area as follows: 99 stations are located in one dense city, while the remaining station is located far away in a small village. Also, assume (the reasonable assumption) that the PM exposure levels measured by the city's stations are extremely correlated, so that for every practical purpose these PM values are the same.

Formally, let y be a 1×100 vector of the PM exposure levels measurments. An environmental study would try to predict y using some model f that is estimated from the exogenous data x, so that $\hat{y} = f(x)$ is the predicted PM values vector. For instance, f might be some linear function of x. Let $\mathcal{F}$ denote the set possible values for f. The quality of the model f can be measured by its loss $\mathcal{L}(f)$, where $\mathcal{L}(f) : \mathcal{F} \to \mathbb{R}$ is called the *Loss function*.

We would like to consider two different loss functions that might be used in the environmental study, and discuss their effect on epidemiological outcomes. The first, is the (unweighted) *quadratic loss function*:

$$\mathcal{L}_{UW} = \Big(y - f(x)\Big)' \Big(y - f(x)\Big), \tag{1}$$

and the second, is the *weighted quadratic loss function*:

$$\mathcal{L}_W = \big(y - f(x)\big)' \mathbf{W} \big(y - f(x)\big). \tag{2}$$

One option is to set the weights matrix $\mathbf{W}$ as the *precision matrix*, i.e. the inverse of the data variance-covariane matrix Σ_n^{-1}. In this case, $\mathcal{L}_W$ can be thought as the squared *Mahalanobis length* of the residuals vector.

define f_{UW} and f_W as the minimizers of $\mathcal{L}_{UW}$ and $\mathcal{L}_W$ over $\mathcal{F}$ respectively. Note that f_{UW} was chosen in a manner that gives each squared error (deviation of y from f_{UW}) an equal weight. Therefore, de facto f_{UW} devotes all of its efforts to predict the values of y in the city, so that the error weight of the the village PM monitoring measurment is practically zero. Without going to much into details, the estimation with f_{UW} would result in poor prediction for the village PM exposure level. Intuitively, f_{UW} does not express the real value of the data since it exaggerates the importance of the 99 city stations while geographically they are actually equivalent to only one observation.

Conversely, with f_W the correlation between observations is taken into account, so that it recognizes that the numerous PM observations in the city are in fact the same observation. Hence the weight of the error in the village is considerable, and so the PM prediction in the village would be more accurate.

Easy to see that under $\mathcal{L}_{UW}$ the model tend to be more acurate in locations where there are many similar observations (which will usually be more populated places) but in the price of a higher error in locations with fewer observations (usually unpopulated remote areas). Also, $\mathcal{L}_W$ is forcing the model to utilize the genuine value that data provide (at least in a geographical sense), hence, predictions accuracy for uncorrelated areas would increase. Moreover, we expect that $\mathcal{L}_{UW}(f_{UW}) < \mathcal{L}_W(f_W)$ since that the error in the village is almost unnoticeable in $\mathcal{L}_{UW}$ relative to $\mathcal{L}_W$.

So which loss function to choose? It depends on what the goal is. A typical epidemiologic study would compare some health indices in places experiencing different pollution levels. For instance, assume that the epidemiologist examines the morbidity rates in the city and the village. He use the average values of the predicted PM levels $\hat{y}$ in these locations as explanatory variables.

Notice that the experimental unit in the epidemiological regression is a geograpical location. Thus, in this example the weight of city and village observations in the epidemiological regression is equal. Clearly, the epidemiologist does not wish that the environmental study would give up on village accuracy for achieving high city accuracy, since that both locations have the same weight in the epidemiological regression. Instead, he prefer that the data would be aqurate in the village just as it is accurate in the city. In other words, from an epidemiological point of view, $\mathcal{L}_W$ is preferred as the the loss function.

However, by examining f with $\mathcal{L}_{UW}$ at the environmental stage, the evaluated model performance ignore the epidemiological research purposes. Moreover, the performance would be over-optimistic from epidemiological perspective.

More generally, we argu that the loss function which evaluates model performance should be

8

constructed so that the regression errors will receive their weights in accordance with the study ultimate goal.

5.1.2 The Folding

Conventional CV techniques assume that experimental units are independent. In data with temporal and spatial observations which are widely common in environmental and geographical studies this assumption is violated. Dependency of observations means that randomly splitted CV (i.e. *K-fold* CV) divides the data into dependent training and validation samples, resulting in overfitting (Larimore and Mehra 1985). In other words, the estimated errors are downward biassed, so that performance estimates are actually overoptimistic (Mosteller and Tukey 1977).

The procedure of CV under dependency conditions has been studied extensively over the last few decade in several contexts. Much progress has been made in the field of nonparametric regression. Hart and Wehrly (1986) for example, proved that when data are positively correlated, using standard CV will overfits for choosing the bandwidth of a kernel estimator in regression. Chu and Marron (1991) proposed a *Modified CV* when selecting the nuisance parameter in nonparametric curve estimation with dependent data. Burman, Chow, and Nolan (1994) continued this line, introducing the *h-block* CV as a verssion of *leav-one-out* (LOO) CV method optimized for use with dependent observations. Their idea is simple: Rather than remove a single case in each CV iteration, remove as well a block of *h* cases from either side of it. they suggest to take *h* as a fixed function of the number of cases, and to correct for the underuse of the sample by adding a term to the estimates.

However, as a verssion of LOO, *h-block* CV has proven to be asymptotically inconsistent (Shao 1993). Racine (2000) proposed the *hv-block* as a modification which is also asymptotically optimal. It extends *h-block* by defining the test set to be *v*-sized cases block instead of being a singletone, while maintaining near-independence of the training and validation data via h-blocking.

Generally, the main CV approach used in the literature to overcome time series dependence, is to choose training set I^t, and validation set I^v such that:

$$\arg\min_{i \in I^t, j \in I^v} |i - j| > h > 0, \tag{3}$$

where *h* expresses the distance (in terms of time) from which observations *i* and *j* are independent. However, these methods were not suitable for data with spatial dependence, and some adaptations were required.

Apperently, CV approaches appropriate for spatialy dependent data recived less attention in the statistical literature. However, some progress has been made in recent years, mainly in the fields of geographical and environmental studies.

R. Telford and Birks (2009) suggested the *Spatial-LOO*, in which the scheme of h-block is adopted for the spatial case by omitting obserations within a radius of *h* from the test set.

They proposed using the range of a variogram model to appropriately define *h*. Trachsel and Telford (2016) proposed further methods for determining *h* e.g. by finding the distance at which the root mean squared error (RMSE) of h-block CV and the RMSE of an independent validation set are similar. Le Rest et al. (2014) considered a variable selection model under conditions of spatial autocorrelated data. They compared a spatial-LOO version with a classical model selection with *Akaike information criterion* (AIC) while accounting for *residual spatial autocorrelation* (RSA). Using simulations they found that spatial-LOO is particulary more useful when the range of RSA was small.

Roberts et al. (2017) examine the utility of blocking procedures for CV in a number of dependent settings, and propose several blocking schemes. They also discuss the possibility of extrapolation (i.e. when blocking hold out entire portion of the predictor space) and state that when extrapolation is the modelling goal deliberate blocking in predictor space should be considered.

Some contribution to blocking approaches comes also from the bootstrap practice. In fact, the procedure of resampling in bootstrap and in CV methods are the same, excepting of with or without replacing. Davison and Hinkley (1997) review some of the schemes for block resampling proposed for complex dependence such as time series and point processes. Examples are: *Post-blackening*, *Blocks-of-blocks*, and *Stationary bootstrap*. They also provide a detailed instruction for finding the optimal of block length under suitable assumptions in an iterative fashion.

5.2 Statistical Model

Although dependency among observations is a common phenomenon in observational studies, it violates one of the standard statistical assumptions, and challenges many classical statistical techniques. Geographical and environmental data generally show both spatial and temporal observations dependency, known as autocorrelation. Standard regression techniques which ignor this dependency structure usually lead to unefficient estimates and bad predictions (N. A. Cressie 1993). Also, prediction errors may be spatially and temporally correlated (Anselin 1998), so that epidemiological studies that use predictions as covariates would suffer from measurment errors (Fuller 2009), resulting in biassed results and erroneous conclusions (Gryparis et al. 2008).

In the last decades there has been an explosion of research in dependence data modeling. *Logitudial* modeling is perhaps the most familiar. It mostly focuses on the dependence among obsevations over time series, and is commonly used in many fields such as biology, economics and more. Several modeling approaches for longitudinal data (also known as panel data in the econometric literature) have been proposed. A very common statistical approach is the use of parameterized covariance models. These models assume an underlying structure of stationary stochastic process, defined by small number of parameters. Such as structures are *Autoregressive* (AR), *Moving Average* (MA) and more. A comprehensive logitudial modeling review is given by Weiss (2005).

Spatial statistics deals with more complex dependency structure, where data represent

observations that are associated with points or regions. According to the first law of geograpy (Tobler 1970), samples in geographical space tend to be more similar, resulting in *spatial autocorrelation* (SAC). Modeling SAC is not a simple task, and is definitely not a straightforward extension of time series into two dimensions.

N. A. Cressie (1993) introdue the *Conditionally Specified Gaussian* model that use the spatial locations of samples to model the probability distribution of the errors under gaussian assumption. The spatial econometrics literature uses the so-called spatial weights matrix W to denote variables lagged in space. W describes the spatial arrangement of the geographical units and can be assumed or estimated. In his seminal book, Anselin (1998) present several estimation methods for linear regression model with spatially dependent error terms from an econometric perspective. LeSage (2008) called this model *Spatial Error Model* (SEM) and considered it as a special case of the more general *Spatial Durbin Model* (SDM). Another approach that has recently proven to be effective is *Bayesian Mierarchical Modeling* (Banerjee, Carlin, and Gelfand 2014), which handles complex relationships such as multi-level data by estimating the parameters of the posterior distribution using bayesian methods.

Spatio-temporal data enable the researcher to take advantage of time-sapce interactions, thus to provide more accurate predictions in a higher resolution. However, the analysis of spatio-temporal data might be quite complicated since both spatial and temporal dependencies should be accounted. It is an emerging reseach field and modeling approaches are still developing. In the environmental exposure assessment studies, Mixed models (Henderson et al. (1959); Robinson (1991)) are probably the most prevalent statistical framework for spatio-temporal data, particularly those analysing satellite based data. Mixed models cope with clustered data by distinguishing between two sources of variation: between clusters, and within clusters.

Another suitable approach for modeling complex dependency structures is the *Generalized Least Squares* (GLS). The GLS (see (Kariya and Kurata 2004) for a comprehensive review) is a linear regression model which uses a variance-covariance matrix of the error terms to efficiently estimate model's parameters in the presence of dependency. The error covariance matrix can be estimated when it is unknown. Since it is a linear model, the GLS has some nice properties as we will discuss later. Altought GLS based models have been known in the statistical literature for decades, their application in geographical and environmental studies has been very limited so far.

In the following we discuss about some of the features of the Mixed model and the GLS, including their estimation and regularization. As we shall see, these two approaches eventually deal with the same challenge: To characterize the observations dependency structure. More specifically, they ask: "how does the residuals variance-covariance matrix is look like?".We state that, as far as regressions are concerened, the almost only difference between models is the definition of the dependence structure through the covariance matrix of the residuals terms.

The GLS model estimate the parameters using a prespecified covariance matrix of the error terms. Thus, it can be thought as a more general approach to model complex structured data, since any covariance matrix can be used. Therefore, the mixed model should be considered

as one of GLS's special cases. Moreover, we argu that any other regression model fall under the extensive settings of the GLS.

Since it is the errors covariance matrix that practically determine the model, we will make an effort to explore its modeling. The estimation of this matrix is undoubtedly a gentle art, as the dependence structure consists of temporal and spatial correlations which their patterns are unknown.

Moreover, whereas the purpose of the model is predication rather than inference, we might consider reducing prediction error by allowing a little bias, with regularization. Suitable regularization approaches for regressions are Shrinkage methods such as *Ridge regression* or *Lasso*. Yet, while such procedures are relatively easy to implement in linear models, it might be more complex in mixed effect models.

5.2.1 Mixed Models

Mixed models, sometimes referred as Hirarchial models, are a class of statistical models suited for the analysis of structured data. Mixed models are particularly useful when obseving repeated measurements of the same statistical units. The mixed models are widely used in environmental studies due to their ability to genuinely combine the data by intruducing multilevel random effects. In theses studies levels are usually time periods, spatial areas, or their interactions.

For each level, the mixed effect model defines clusters. In time level, typical clusters are days or hours, and in spatial level, they might be grid cells. The model assumes that observations between clusters are independent, while obserrvations within cluster are dependent since they belong to the same subpopulation. For instance, when days are the only clusters in a PM spatio-temporal dataset, PM measurments for a specific day of all geographic units are dependent, as they are assumed to be drawn from the same subpopulation. That is to say, that every day is unique in its distribution of PM measurments across different geographic locations. This cluster-specific uniqueness is reflected in a estimated posteriori coefficient and reffered as *random effects*. Other model coefficients are fixed across clusters (usually referred as *fixed effects*) and have the same meaning as in standard regression models.

A very common model in the exposure asseement literatue is the *Linear Mixed Effect* (LME) model that was originally developed by Laird and Ware (1982). It can be formalated as:

$$y_j = X_j\beta + Z_j b_j + \varepsilon_j \qquad j = 1, ..., T \qquad (4)$$

where:

j represent a cluster, s_j is the number of observations in cluster j, T is the number of clusters and $N = \sum_{j=1}^{T} s_j$; y_j is an $s_j \times 1$ vector of responses of the jth cluster; X_j is a $s_j \times m$ design matrix of fixed effects; β is an $m \times 1$ fixed effects coefficients; Z_j is an $s_j \times k$ design matrix of random effects; b_j is an $k \times 1$ random effects coefficients with mean zero and covariance

matrix $\sigma^2 D$; and ε_j is an $s_j \times 1$ independent and identically distributed (*iid*) error terms vector. Each element in ε_j is assumed to have mean zero and variance σ^2.

The compressed matrix form of N equations is:

$$y = X\beta + Zb + \varepsilon, \tag{5}$$

or,

$$y = X\beta + \eta, \tag{6}$$

where y and η are $N \times 1$ vectors and X is an $m \times N$ matrix. Notice that:

$$\eta = \begin{bmatrix} \eta_1 \\ \vdots \\ \eta_T \end{bmatrix} = \begin{bmatrix} \varepsilon_1 + Z_1 b_1 \\ \vdots \\ \varepsilon_T + Z_T b_T \end{bmatrix}. \tag{7}$$

The model assumes that $E(\eta) = 0$. Note that $\text{Var}(\eta)$ is an $N \times N$ covariance matrix. Let us define $V = \text{Var}(\eta) = E(\eta\eta')$. V has the following block diagonal form:

$$V_{N\times N} = \sigma^2 \begin{bmatrix} I_{s_1} + Z_1 D Z_1' & 0 & 0 & 0 \\ 0 & I_{s_2} + Z_2 D Z_2' & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & I_{s_T} + Z_T D Z_T' \end{bmatrix}, \tag{8}$$

where I_{s_j} is the identity matrix of size s_j.

In other words, without further assumptions (e.g. residuals autocorrelation), the LME can be considered as the familiar linear model, exept that it assumes that the residuals covariance matrix V follows a specific structure. In particular, V is set to have a diagonal block design, where each block represents a cluster.

Note that as more levels are added (i.e. more clusters), the less sparse the covariance matrix would be. However, block design covariance matrix does not allow for correlation between clusters. Lack of correlation between clusters is very unlikely when the clusters are spatial or time units, therefore some adjustment could be useful.

5.2.2 GLS

The GLS (first described by Aitken (1936)), extends the Gauss–Markov theorem to the case where the covariance of the error terms is not a scalar matrix.

To understand GLS estimator, consider the linear regression model in (6), but now, without any assumptions about the error term:

$$y = X\beta + \varepsilon. \tag{9}$$

According to Gauss-Markov theorem, for a known covariance matrix of the error terms Σ, the best linear unbiased estimator (BLUP) for β is:

$$b(\Sigma) = (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}y. \tag{10}$$

However, Σ is usually unknown, and so *GLS estimators* (GLSE) replace Σ with the its estimated value:

$$\hat{\beta}_{GLS} = b(\hat{\Sigma}) = (X'\hat{\Sigma}^{-1}X)^{-1}X'\hat{\Sigma}^{-1}y. \tag{11}$$

Clearly, the quality of GLSE lies in the estimation of Σ. In the proposed study, we will examine and discuss several estimation methods of Σ. Now, it is easy to realize that GLSE includes as its special cases various specific models that are determined by the estimated error covariance matrix.

We also would like to emphasize that GLS estimator is the minimizer of the squared Mahalanobis length of the residual vector $y - X\beta$. This is particulary important since we may want to choose an estimator that minimizes the aforementioned weighted loss function with estimated Σ^{-1} as the weight matrix:

$$\hat{\beta}_{GLS} = \arg\min_{\beta}\left\{(y - X\beta)^{-1}\hat{\Sigma}^{-1}(y - X\beta)\right\}. \tag{12}$$

The GLSE can be considered as an estimation method that de-correlate the scale of the *ordinary least squares* (OLS) errors. This means that as long as we reasonably estimate Σ, strongly dependent observations, which usually have highly correlated errors, would have less impact on the estimator values than independent observations.

5.2.3   The Errors Variance-covariance Matrix

Whether it's a Mixed model, AR, or SEM, it is the covariance matrix that essentially tells the story of data dependency. The decision regarding the errors covariance modeling is the researcher's statement about the data generating process.

Here we review several models specifications which may be applied. We focus on *parameterized* covariance matrices, where all the components of the covariance matrix are a function of $q \in \{1, ..., N(N+1)/2\}$ parameters, where N is the number of error terms. As an initial step, our concentration will be in *stationary covariance functions*. According to N. Cressie and Wikle (2015), f is a stationary spatio-temporal covariance function on $\mathbb{R}^d \times \mathbb{R}$ (where d is the spatial dimension) if it can be written as:

$$f((s;t),(x;r)) = C(s-x;t-r), \qquad s,x \in \mathbb{R}^d, \quad t,r \in \mathbb{R}. \tag{13}$$

We define the covariance model as $\Sigma_C(\theta)$, where C indicates the selected covariance function and θ the distinct $q \times 1$ parameters vector that can be estimated from the data. Weiss (2005) states that the covariance model $\Sigma_C(\theta)$ should be chosen so that the true covariance function C is from the type defined be the model (i.e. AR, MA, etc.), but parsimonious, i.e. q is small as possible.

We point that any parameterized covariance matrix $\Sigma_C(\theta)$ can be considered as a compromise between two possibilities:

$$\Sigma_s(\theta) = \sigma^2 \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & & \\ \vdots & & \ddots & \\ 0 & & \dots & 1 \end{bmatrix} \quad \text{and} \quad \Sigma_u(\theta) = \begin{bmatrix} \sigma_1^2 & \sigma_{1,2} & \dots & \sigma_{1,n} \\ \sigma_{2,1} & \sigma_2^2 & & \\ \vdots & & \ddots & \vdots \\ \sigma_{n,1} & & \dots & \sigma_n^2 \end{bmatrix},$$

where $\Sigma_s(\theta)$ is the variance-scalled identity matrix, somtimes called *spherical error variance* matrix (Hayashi 2000). This structure assumes *homoscedasticity* and no autocorrelation, and requires estimation of only one parameter. $\Sigma_u(\theta)$ on the other hand, is the most general model, called the *unstructured covariance matrix* and specifies no patterns. Unfortunately, the use of unstructured covariance matrix is not feasible in most cases since it requires fitting $N(N+1)/2$ parameters. This requires high number of measurements for every time-space interaction unit to achieve nonsingularity, which most datasets do not support.

We would like to examine covariance models at an increasing complexity. Firstly, we will separately discuss temporal and spatial modeling. Further we will examine a space-time integrated model. We stick to a spatio-temporal framework in which $i \in 1, ..., S$ indicates a spatial unit and $j \in 1, ..., T$ indicates a time unit.

5.2.3.1 Temporal perspective

When PM measurments are regressed agisnt environmental covariates, both the response and predictors vary over time. Thus a case to suspect is errors autocorrelation. A common approach to describe the errors covariance matrix of a process like this is the errors *autoregressive* (AR) model. We illustrate it by considering the errors AR(1) model, in which the error term depends on its (1) previous values. This model can be readily extended to AR(p).

Consider the model in (9), only with the following extension: A component in the $N \times 1$ vector ε which corresponds to the error of spatal unit i and time unit j can be written as:

$$\varepsilon_{ij} = \rho\varepsilon_{i(j-1)} + \delta_{ij}, \tag{14}$$

where, ρ is referred as the *autocorrelation* parameter, and δ_{ij} is white noise iid process and follows a normal distribution: $\delta_{ij} \sim \mathcal{N}(0, \sigma_\delta^2)$. Notice that the process is defined as *wide-sence stationary* when $|\rho| < 1$ (Weiss 2005). In this case the correlation function would be:

$$\text{Corr}(\varepsilon_{ij}, \varepsilon_{il}) = \rho^{|j-l|}, \tag{15}$$

We ignore here the spatial pattern and assume that the process is spatially constant. That is, the $N \times N$ covariance matrix has the form:

$$\Sigma = I_S \otimes \tau^2 \begin{bmatrix} 1 & \rho & \rho^2 & & \rho^{T-1} \\ \rho & 1 & \rho & \cdots & \rho^{T-2} \\ \rho^2 & \rho & 1 & & \rho^{T-3} \\ & \vdots & & \ddots & \vdots \\ \rho^{T-1} & \rho^{T-2} & \rho^{T-3} & \cdots & 1 \end{bmatrix}, \tag{16}$$

where I_S is the $S \times S$ identity matrix, $\otimes$ is the *Kronecker product*, and $\tau^2 = \text{Var}(\varepsilon_{ij}) = \frac{\sigma_\delta^2}{1-\rho^2}$. Note that this AR(1) covariance model requires the estimation of 2 parameters, and in general $p+1$ paramrters are required to specify an AR(p) covariance model.

There are many more alternatives for the errors temporal covariance model. Another example which is also more general than the AR model, is a *Toeplitz* covariance matrix (see Schott (2016)). The complete $n \times N$ matrix is then defined as:

$$\Sigma = I_S \otimes \begin{bmatrix} \sigma_0 & \sigma_1 & \sigma_2 & & \sigma_{T-1} \\ \sigma_1 & \sigma_0 & \sigma_1 & \cdots & \sigma_{T-2} \\ \sigma_2 & \sigma_1 & \sigma_0 & & \sigma_{T-3} \\ & \vdots & & \ddots & \vdots \\ \sigma_{T-1} & \sigma_{T-2} & \sigma_{T-3} & \cdots & \sigma_0 \end{bmatrix}. \tag{17}$$

The Toeplitz covariance matrix requires the estimation of T parameters.

5.2.3.2 Spatial perspective

There are several approaches discussed in the literature regarding the covariance model in SAC data, each implies a different assumption about the spatial pattern. One approach, which is particularly common in econometric studies, is to model the errors generating process through a *Weight matrix* - W (first introduced by Ord (1975), but see also Elhorst (n.d.)). This can be described as follows: Consider the model in (9), but now with different assumption regarding to the $i - j$th element of ε:

$$\varepsilon_{ij} = \lambda \sum_{k=1}^{S} W_{ik}\varepsilon_k + \nu_i \tag{18}$$

where W_{ik} is the $i - k$th element in the $S \times S$ weight matrix W, and ν_i is an iid white noise with $\text{E}(\nu_i) = 0$ and $\text{Var}(\nu_i) = \sigma_\nu^2$. Note this scheme ignore temporal effects by assuming the same structure for each time unit j. $\dot{\varepsilon}_{ij} = \sum_{k=1}^{S} w_{ik}\varepsilon_{kj}$ is called the *spatial lag*, since it

represent a linear combination of (spatialy) neighboring errors values. λ is the correlation bettween the errors and their spatial lags. In matrix notation for each time unit j, the $S \times 1$ vector ε can be represented as:

$$\varepsilon = (I - \lambda W)^{-1}\nu. \tag{19}$$

Note that $E(\varepsilon) = 0$ and its variance is:

$$\Sigma = I_T \otimes \sigma_\nu^2 (I - \lambda W)^{-1}(I - \lambda W')^{-1} \tag{20}$$

Plenty of alternatives for choosing the components W_{ik} of W exist, some of the most common are:

- *K-nearest neighbors*:

$$W_{ik} = \begin{cases} 1 & , i \in N_K(k) \\ 0 & , otherwise \end{cases} \tag{21}$$

- *Radial Distance*:

$$W_{ik} = \begin{cases} 1 & , 0 \le d_{ik} \le L \\ 0 & , otherwise \end{cases} \tag{22}$$

- *Power Distance*:

$$W_{ik} = \frac{1}{d_{ik}^\alpha} \tag{23}$$

Anoter approach to model the covariance is to specifiy directly the spatial correlation matrix components K_{ik}, assuming a functional form. After constructing the correlation matrix K taking $\Sigma = I_T \otimes \sigma_\varepsilon^2 K$ will spacify the complete covariance matrix. Some instances for the K_{ik} function are:

- *Negative exponential*:

$$K_{ik} = b_1 \exp(-\frac{d_{ik}^\alpha}{b_2}) \tag{24}$$

Note that when $\alpha = 2$ the negative exponential is exactly a *gaussian*.

- *Spherical*:

$$K_{ik} = \begin{cases} b_1(1 - \frac{3d_{ik}}{2b_2} + \frac{d_{ik}^3}{2b_2^3}) & , 0 \le d_{ik} < b_2 \\ 0 & , d_{ik} > b_2 \end{cases} \tag{25}$$

The *variogram* (see N. A. Cressie (1993)) is another approach that can be used to descibe a spatial dependence in a stochastic process. It is very popular in the domain of geostatistics,

as it is used in *kriging* technique (see for example Stein (2012)). A stationary variogram 2γ of a spatial process $\varepsilon(s) : s \in D_s \subset \mathbb{R}^d$ is defined as the variance of the difference h between two field values (here values of errors) at spatial locations s and $s + h$:

$$2\gamma(h) = \text{Var}(\varepsilon(s + h) - \varepsilon(s)), \qquad \text{for all } s, s + h \in D_s \tag{26}$$

where γ is called the *semivariogram*. If the process is furthermore *isotropic*, then the variogram can be described as a function of $\|h\|$. After constructing the variogram function, the covariance matrix is readily defined.

5.2.3.3 Spatio-temporal perspective

The estimation of a spatio-temporal error covariance model is a complex task. We will start by examining a relatively simple model by assuming a *separable* spatio-temporal covariance function. That is, for $s, x \in \mathbb{R}^d$ and $t, r \in \mathbb{R}$

$$\text{Cov}\big(\varepsilon(s; t), \varepsilon(x; r)\big) = C^{(s)}(s, x) \cdot C^{(t)}(t, r) , \tag{27}$$

where $C^{(s)}$ and $C^{(t)}$ are the spatial and temporal cpvariance function. Under spatio-temporal separability, the covariance matrix can be written as a Kronecker product of the separatly estimated spatial and a temporal matrices (Huizenga et al. (2002); Genton (2007)). The main reason to choose a separable covariances structure is due to the reduction in the number of estimated parameters, and a significantly decrease in computational complexity. When the dataset is large in relation to calculation capabilities, this can be a worthwhile choice.

However, the separable covariance model class is limited since it does not account for space-time interaction. N. Cressie and Huang (1999) give some methodology for developing whole classes of *nonseparable* spatio-temporal stationary covariance functions in closed form. Also, a more recent review is provided by N. Cressie and Wikle (2015). They discuss in details nonseparable covariance as well as variogram models, including examples and visualisations.

Except for the functional form, another fundamental issue, is the fact that in contrast to other spatio-temporal variables, we do not actually observe the error terms. OLS residuals are frequently used as empirical error terms, sometimes as an initial stage in an iterative procedure (see Kariya and Kurata (2004); Fomby, Hill, and Johnson (2012)). In this case the error covariance matrix is quite sensitive to the OLS regression. Hence, it is important that the OLS regression residuals reflect the true functional structure of the errors.

5.2.4 Estimation Perspective

5.2.4.1 Estimation of the Mixed Model

Back to the simple LME model. The mixed effect model assume normal disttribution of the error, specificaly:

$$\varepsilon_j \sim \mathcal{N}(0, \sigma^2 I) \qquad b_j \sim \mathcal{N}(0, \sigma^2 D), \tag{28}$$

where $I = I_{s_j}$.

The multivariate normal distribution of y_j can then be writen as:

$$y_j \sim \mathcal{N}(X_j \beta, \sigma^2 (I + Z_j D Z_j')), \tag{29}$$

and the log likelihood function for the linear mixed model is given by:

$$l(\beta, \sigma^2, D) = -\frac{N}{2} \ln 2\pi - \frac{1}{2} \left(N \ln \sigma^2 + \sum_{j=1}^{N} \left(\ln |I + Z_j D Z_j'| + \sigma^2 (y_j - X_j \beta)' (I + Z_j D Z_j')^{-1} (y_j - X_j \beta) \right) \right) \tag{30}$$

This log likelihood function involves matrices inverse and determinante which might be a difficult task when the matrices size is large. However, some dimension reduction formulation can be employed in order to make calculation easier, see for instance Demidenko (2013).

5.2.4.2 Estimation of the GLS

When Σ is known, GLS estimation is essentially applying OLS to the transformed data. To see this, consider Σ's Cholesky's decomposition: $\Sigma = L\Lambda L'$ where L is a unitriangular matrix and Λ is a diagonal matrix. Easy to see that:

$$\Sigma^{-1} = PP', \tag{31}$$

where $P = L^{-1}\Lambda^{-\frac{1}{2}}$ and $P\Sigma P' = I$.

Notice that multiplying both sides of (9) by P yields:

$$\tilde{y} = \tilde{X}\beta + \tilde{\varepsilon}, \tag{32}$$

where, $\tilde{y} = Py$, $\tilde{X} = PX$ and $\tilde{\varepsilon} = P\varepsilon$. Also note that $E(\tilde{\varepsilon}) = 0$ and $\text{Var}(\tilde{\varepsilon}) = E(P\varepsilon\varepsilon'P') = \sigma^2 P\Sigma P' = \sigma^2 I$, hence the GLS estimator β_{GLS} is achieved by minimizing the sum of the squares (i.e. apply OLS) of (32):

$$\begin{aligned} \hat{\beta}_{GLS} &= (\tilde{X}'\tilde{X})^{-1}\tilde{X}'\tilde{y} \\ &= (X'P'PX)^{-1}X'P'Py \\ &= (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}y. \end{aligned} \tag{33}$$

However, Σ is usually unknown and need to be estimated. The estimators in this case are sometimes called *feasible generalized least squares* (FGLS). Asymptotically, under appropriate

conditions, all properties of FGLS are common with respect to GLS (Fomby, Hill, and Johnson 2012). The FGLS estimation proceeds in two stages which can be repeated for several iterations: (i) The model is estimated using (33) and assuming Σ is known. The residuals $e = y - X\beta$ are then used as the empirical errors, to construct the error covariance matrix by estimating θ (see section 5.2.3). (ii) The GLS estimation is performed using the previous stage estimated $\Sigma(\hat{\theta})$. For the first iteration it is usually assumed that there is no dependence structure in the data so that: $\Sigma = I$, i.e. OLS estimation. Note that this procedure is allowed only if the asymptotic covariance between β and θ is zero [ref][].

Note that by assuming normal disstribution of the error term ε, the log-likelihood function of (9) can be writen as:

$$l(\beta, \theta) = -\frac{N}{2} \log(2\pi) - \frac{1}{2} \log |\Sigma(\theta)| - \frac{1}{2}(y - X\beta)'\Sigma(\theta)^{-1}(y - X\beta) \tag{34}$$

Parameters can also be estimated using *Maximum likelihood estimators* (MLE) w.r.t the log-likelihood in (34). Note that (34) is a general form of (30) (to see this, just set $\Sigma(\hat{\theta})$ to the specific diagonal format). We find it important to note that when gaussian distribution of errors is assumed, the MLE β_{ML} is always identical to GLSE of the form $b(\Sigma(\hat{\theta}))$ (Kariya and Kurata 2004). Therefore, another advantage of the GLSE is that it does not requires gaussian or other specific disstribution of the data.

5.2.5 Regularization

Reducing the variance of the predicted values can be done by *shrinking*, while sacrificing a little bit of bias. When the goal is to improve prediction accuracy this should be considered. Shrinkage methods are continuous selection models (not limited to discrete variable selection) that impose a penalty on the regression coefficients. The most familiar are *ridge regression*, the *least absolute shrinkage and selection operator* (Lasso), and the *elastic net*, (see Hastie et al. (2009) , Ch. 3.4 for an enlightening review). However, these methods are mostly suitable for data with independent observations, and are not straightforward in a spatio-temporal datasets.

5.2.5.1 Regularization in the Mixed Model

Altought regularization in regression models have received considerable attention over the past years, literature on regularized LME models is somewhat scarce. The challenge in regularization of mixed models is to properly select random effects together with the fixed effects. This challenge stems from the fact that as long as the random effects are not determined its covariance matrix is unknown. One option is to perform selection in separate stages, but it may lead to different regularization solutions depending on the order of the stages.

Recently, several procedures have been proposed to identify both the random and fixed effects. Bondell, Krishna, and Ghosh (2010) proposed a simultaneous selection of the fixed and

random effects in an LME model using a modified Cholesky decomposition. Their method is based on a penalized joint log-likelihood with an adaptive penalty (*adaptive Lasso*). Y. Fan and Li (2012) proposed to use a proxy matrix in the penalized profile likelihood to overcome the difficulty of unknown covariance matrix of the random effects. One drawback of these kind of methodsis is that they usually involve complex numerical optimization, therefore are computational intensive in relation to classical regularizations methods

5.2.5.2 Regularization in GLS

As described, in GLS estimaton the OLS is implemented on the whitening transformation of the data. Therefore, its regularization formulation is as easy as OLS regularization model:

$$
\begin{aligned}
\hat{\beta}_{RGLS} &= \arg\min_{\beta}\Big\{(y - X\beta)'\hat{\Sigma}^{-1}(y - X\beta) + \lambda g(\beta)\Big\} \\
&= \arg\min_{\beta}\Big\{(\tilde{y} - \tilde{X}\beta)'(\tilde{y} - \tilde{X}\beta) + \lambda g(\beta)\Big\}
\end{aligned}
\tag{35}
$$

where $g(\beta)$ is some penalization on model complexity. For instance by setting: $g(\beta) = \sum_{i=k}^{p} \beta_k^2$ we get the ridge regression estimator:

$$
\hat{\beta}_{RGLS} = \beta_{ridge} = (\tilde{X}'\tilde{X} + \lambda I)^{-1}\tilde{X}'\tilde{y}
\tag{36}
$$

Observe that β_{RGLS} does not involve numeric optimization and enjoys a closed form solution.

5.3 Example: Validation and Estimation from Epidemiological Perspective

EIV.Rmd file goes here

ask johnathan if necessary

5.4 Computational Challenges

The analysis of increasingly large scale data is an active research area in statistics and machine learning. Over the last decade, environmental databases have grown tremendously in terms of voulume, intensity and complexity (Hampton et al. 2013). However, large scale databases pose new barriers, primarily: computer memory and computing power (C. Wang et al. 2015). We propose to take advantage of some the recent methodological and software developments to address these challenges.

5.4.1 Sparse Representations

J. H. Wilkinson defined a sparse matrix to be "any matrix with enough zeros that it pays to take advantage of them" (Björck 1996). Due to its nature as suitable to be efficiently represented, sparse data is more easily compressed and thus require significantly less storage. Efficient representation of data in memory reduce computing time, and allow to fit models that would otherwise require tremendous amounts of memory (Rosenblatt 2017). Moreover, sparse matrices are desirable in scientific largescale computations thanks to *sparse matrix algorithms.* These algorithms take advantage of the sparse structure of the matrix by avoiding arithmetic operations on zero elements. Therefore, operations such as matrix multiplication, inversion and determinant calculation are much faster when matrices are sparse.

Our PM assesment model might enjoy sparse representation in two aspects. First, since that in the statistical software we use - R, explanatory factors are actually converted to numeric vectors with many zeors when fitting a model. Second, our proposed model requires a precision matrix that is very likely to have many zero entries.

In R, we may use the **Matrix** package (Bates and Maechler 2010) which provide data storage classes for sparse matrices. Fitting a model to these classes can be implemented with **MatrixModels** packge (Bates and Maechler, n.d.). Using these packages for implementation of statistical algorithms on sparse class objects may save considerable memory and computing-time and reduce computational burden (see Rosenblatt (2017) for demonstration of this assertion)

Computational challenges demand us to devote much attention to sparse considerations. The main barrier to speedy computation of the estimates in GLS, lies in the $N \times N$ nature of the errors covariance matrix; Thus we may consider to use sparse matrix techniques to facilitate computation in large data (e.g. Pace (1997)). In particular, we might want to estimate the covariance matrix using some regularization-based *thresholding estimation* (J. Fan, Liao, and Liu 2016), so that entries of weakly correlated observations would be zeros. Another option is to chose the functional form of the error covariance matrix so that its inverse would have simple sparse structure. For instance, the inverse of the AR(1) based matrix proposed in (16) has a convenient *band* form:

$$\Sigma^{-1} = I_S \otimes \tau^{-2} \begin{bmatrix} 1 & -\rho & & & 0 \\ -\rho & 1+\rho^2 & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & 1+\rho^2 & -\rho \\ 0 & & & -\rho & 1 \end{bmatrix}, \tag{37}$$

5.4.2 Memory Efficiency

When dataset is large relative to RAM (J. W. Emerson and Kane (2012) suggested to considere a dataset that exceeds 20% of RAM as *large*), computing from RAM might be problematic

even when data is sparse, especially when it comes to advanced statistical algorithms. This problem can be especially significant for R users due to R's in-RAM storage mechanism. To overcome this hurdle, one may use *external memory algorithms* (EMA) which works by storing the data on the local storage (HD, SSD, etc.), and processing one chunk of it at a time in RAM (see Vitter (2001)). In that way, files are very fast accessed since operations are handled at the operating system level. This procedure (sometimes reffered as memory mapping) allows to quickly save and compute directly from local storage.

Currently, wide range of R packages such as **bigmemory** (Kane et al. 2013) and **ff** (Adler et al. 2014), follow this technology and provide data structures for large and massive data that can be approached as R objects. Respectively, **biganalytics** (J. W. Emerson and Kane 2013) and **ffbase** (Jonge, Wijffels, and Laan 2014) provide specific implementation of data functions such as regression and classification models for objects defined by these packages. A comprehensive review by C. Wang et al. (2015) presents more R packages that can help breaking the memory barrier (see also R Archive Network (CRAN) task view (Eddelbuettel 2017)).

5.4.3   Parallel Computing

When the bottleneck is due to CPU and not RAM loads, one encounters a computing power barrier. Breaking computing power barriers can be done by parallelisation, i.e. applying multiple processors to a single task. The idea is as follows: data sets are splitted into "chunks" and then the analysis is performed by multiple machines in parallel (Schmidberger et al. 2009). For independent chuncks this scheme can be seen as so-called *embarrassingly parallel.* When the task involves learning a statistical models (i.e. a machine learning algorithem) this procedure sometimes reffered as *distributed machine learning.* Learning task of each machine might be parameters estimation, prediction or other tasks. The outcome of a distributed learning is obtained by some aggregation procedure of the machines outputs, for example: averaging.

Empirical risk minimization (ERM) algorithems such as linear models can be fastly computed using parallel scheme. Under certain conditions, the obtained estimator is as accurate as the centralized one. Rosenblatt and Nadler (2016) studied the optimality of the averaged ERM estimator and proved that it is first-order equivalence with the centralized estimators in a classical low-dimentional asymptotic settings.

We propose to exploit R's packags which offer a variety of techniques to execute parallel computing (see a review by Chapple et al. (2016)). For example, the **foreach** package (Analytics and Weston 2015) provides a general framework for implamenting parallel algorithems and can exploit the shared memory capabilities of **bigmemory**. It facilitates executing loops in parallel with different parallel mechanisms. These mechanisms determine the form of communication between machines and provided by **multicore**, **parallel**, **Rmpi** and **snow**. Demonstration of the performance of these packages is given in Rosenblatt (2017).

6 Preliminary Results

7   Appendix

References

Adler, D, C Gläser, O Nenadic, J Oehlschlägel, and W Zucchini. 2014. "Ff: Memory-Efficient Storage of Large Data on Disk and Fast Access Functions." *R Package Version*, 2–2.

Aitken, Alexander C. 1936. "IV.—on Least Squares and Linear Combination of Observations." *Proceedings of the Royal Society of Edinburgh* 55. Royal Society of Edinburgh Scotland Foundation: 42–48.

Analytics, Revolution, and Steve Weston. 2015. "Foreach: Provides Foreach Looping Construct for R." *R Package Version* 1 (3): 1.

Anselin, L. 1998. "Spatial Econometrics: Methods and Models." Kluwer Academic Publisher.

Arlot, Sylvain, Alain Celisse, and others. 2010. "A Survey of Cross-Validation Procedures for Model Selection." *Statistics Surveys* 4. The author, under a Creative Commons Attribution License: 40–79.

Banerjee, Sudipto, Bradley P Carlin, and Alan E Gelfand. 2014. *Hierarchical Modeling and Analysis for Spatial Data.* Crc Press.

Bates, Douglas, and Martin Maechler. 2010. "Matrix: Sparse and Dense Matrix Classes and Methods." *R Package Version 0.999375-43, URL Http://Cran. R-Project. Org/Package= Matrix.*

———. n.d. "MatrixModels: Modelling with Sparse and Dense Matrices, 2015." *URL Http://CRAN. R-Project. Org/Package= MatrixModels. R Package Version 0.4-1.*

Björck, Åke. 1996. *Numerical Methods for Least Squares Problems.* SIAM.

Bondell, Howard D, Arun Krishna, and Sujit K Ghosh. 2010. "Joint Variable Selection for Fixed and Random Effects in Linear Mixed-Effects Models." *Biometrics* 66 (4). Wiley Online Library: 1069–77.

Briggs, David J, Cornelis de Hoogh, John Gulliver, John Wills, Paul Elliott, Simon Kingham, and Kirsty Smallbone. 2000. "A Regression-Based Method for Mapping Traffic-Related Air Pollution: Application and Testing in Four Contrasting Urban Environments." *Science of the Total Environment* 253 (1). Elsevier: 151–67.

Burman, Prabir, Edmond Chow, and Deborah Nolan. 1994. "A Cross-Validatory Method for Dependent Data." *Biometrika* 81 (2). Oxford University Press: 351–58.

Chapple, Simon R, Eilidh Troup, Thorsten Forster, and Terence Sloan. 2016. *Mastering Parallel Programming with R.* Packt Publishing Ltd.

Chu, C-K, and James S Marron. 1991. "Choosing a Kernel Regression Estimator." *Statistical Science.* JSTOR, 404–19.

Chudnovsky, Alexandra A, Petros Koutrakis, Itai Kloog, Steven Melly, Francesco Nordio, Alexei Lyapustin, Yujie Wang, and Joel Schwartz. 2014. "Fine Particulate Matter Predictions

Using High Resolution Aerosol Optical Depth (Aod) Retrievals." *Atmospheric Environment* 89. Elsevier: 189–98.

Cressie, Noel AC. 1993. "Statistics for Spatial Data." Wiley.

Cressie, Noel, and Hsin-Cheng Huang. 1999. "Classes of Nonseparable, Spatio-Temporal Stationary Covariance Functions." *Journal of the American Statistical Association* 94 (448). Taylor & Francis Group: 1330–9.

Cressie, Noel, and Christopher K Wikle. 2015. *Statistics for Spatio-Temporal Data.* John Wiley & Sons.

Davison, Anthony Christopher, and David Victor Hinkley. 1997. *Bootstrap Methods and Their Application.* Vol. 1. Cambridge university press.

Demidenko, Eugene. 2013. *Mixed Models: Theory and Applications with R.* John Wiley & Sons.

Eddelbuettel, Dirk. 2017. "CRAN Task View: High-Performance and Parallel Computing with R."

Elhorst, J Paul. n.d. *Spatial Econometrics: From Cross-Sectional Data to Spatial Panels.* Springer.

Emerson, John W, and Michael J Kane. 2012. "Don't Drown in the Data." *Significance* 9 (4). Wiley Online Library: 38–39.

————. 2013. "Biganalytics: A Library of Utilities for Big. Matrix Objects of Package Bigmemory." *R Package Version* 1 (1).

Fan, Jianqing, Yuan Liao, and Han Liu. 2016. "An Overview of the Estimation of Large Covariance and Precision Matrices." *The Econometrics Journal* 19 (1). Wiley Online Library.

Fan, Yingying, and Runze Li. 2012. "Variable Selection in Linear Mixed Effects Models." *Annals of Statistics* 40 (4). NIH Public Access: 2043.

Fomby, Thomas B, R Carter Hill, and Stanley R Johnson. 2012. *Advanced Econometric Methods.* Springer Science & Business Media.

Fuller, Wayne A. 2009. *Measurement Error Models.* Vol. 305. John Wiley & Sons.

Genton, Marc G. 2007. "Separable Approximations of Space-Time Covariance Matrices." *Environmetrics* 18 (7). Wiley Online Library: 681–95.

Gryparis, Alexandros, Christopher J Paciorek, Ariana Zeka, Joel Schwartz, and Brent A Coull. 2008. "Measurement Error Caused by Spatial Misalignment in Environmental Epidemiology." *Biostatistics* 10 (2). Oxford University Press: 258–74.

Hampton, Stephanie E, Carly A Strasser, Joshua J Tewksbury, Wendy K Gram, Amber E Budden, Archer L Batcheller, Clifford S Duke, and John H Porter. 2013. "Big Data and the Future of Ecology." *Frontiers in Ecology and the Environment* 11 (3). Wiley Online Library:

156–62.

Hart, Jeffrey D, and Thomas E Wehrly. 1986. "Kernel Regression Estimation Using Repeated Measurements Data." *Journal of the American Statistical Association* 81 (396). Taylor & Francis Group: 1080–8.

Hastie, Trevor, Robert Tibshirani, JH Friedman, and others. 2009. "The Elements of Statistical Learning." Springer,

Hayashi, Fumio. 2000. "Econometrics." Princeton University Press Princeton.

Henderson, Charles R, Oscar Kempthorne, Shayle R Searle, and CM Von Krosigk. 1959. "The Estimation of Environmental and Genetic Trends from Records Subject to Culling." *Biometrics* 15 (2). JSTOR: 192–218.

Huizenga, Hilde M, Jan C De Munck, Lourens J Waldorp, and Raoul PPP Grasman. 2002. "Spatiotemporal Eeg/Meg Source Analysis Based on a Parametric Noise Covariance Model." *IEEE Transactions on Biomedical Engineering* 49 (6). IEEE: 533–39.

Jonge, Edwin de, Jan Wijffels, and Jan van der Laan. 2014. "Ffbase: Basic Statistical Functions for Package Ff. R Package Version 0.11. 3."

Kane, Michael J, John Emerson, Stephen Weston, and others. 2013. "Scalable Strategies for Computing with Massive Data." *Journal of Statistical Software* 55 (14): 1–19.

Kariya, Takeaki, and Hiroshi Kurata. 2004. *Generalized Least Squares.* John Wiley & Sons.

Kloog, Itai, Petros Koutrakis, Brent A Coull, Hyung Joo Lee, and Joel Schwartz. 2011. "Assessing Temporally and Spatially Resolved Pm 2.5 Exposures for Epidemiological Studies Using Satellite Aerosol Optical Depth Measurements." *Atmospheric Environment* 45 (35). Elsevier: 6267–75.

Kloog, Itai, Steven J Melly, William L Ridgway, Brent A Coull, and Joel Schwartz. 2012. "Using New Satellite Based Exposure Methods to Study the Association Between Pregnancy Pm 2.5 Exposure, Premature Birth and Birth Weight in Massachusetts." *Environmental Health* 11 (1). BioMed Central: 40.

Kloog, Itai, Meytar Sorek-Hamer, Alexei Lyapustin, Brent Coull, Yujie Wang, Allan C Just, Joel Schwartz, and David M Broday. 2015. "Estimating Daily Pm 2.5 and Pm 10 Across the Complex Geo-Climate Region of Israel Using Maiac Satellite-Based Aod Data." *Atmospheric Environment* 122. Elsevier: 409–16.

Laird, Nan M, and James H Ware. 1982. "Random-Effects Models for Longitudinal Data." *Biometrics.* JSTOR, 963–74.

Larimore, Wallace E, and Raman K Mehra. 1985. "Problem of Overfitting Data." *Byte* 10 (10): 167–78.

Le Rest, Kévin, David Pinaud, Pascal Monestiez, Joël Chadoeuf, and Vincent Bretagnolle. 2014. "Spatial Leave-One-Out Cross-Validation for Variable Selection in the Presence of

Spatial Autocorrelation." *Global Ecology and Biogeography* 23 (7). Wiley Online Library: 811–20.

Lee, Mihye, Itai Kloog, Alexandra Chudnovsky, Alexei Lyapustin, Yujie Wang, Steven Melly, Brent Coull, Petros Koutrakis, and Joel Schwartz. 2016. "Spatiotemporal Prediction of Fine Particulate Matter Using High-Resolution Satellite Images in the Southeastern Us 2003–2011." *Journal of Exposure Science and Environmental Epidemiology* 26 (4). Nature Publishing Group: 377–84.

LeSage, James P. 2008. "An Introduction to Spatial Econometrics." *Revue d'économie Industrielle*, no. 3. De Boeck Supérieur: 19–44.

Madrigano, Jaime, Itai Kloog, Robert Goldberg, Brent A Coull, Murray A Mittleman, and Joel Schwartz. 2013. "Long-Term Exposure to Pm2. 5 and Incidence of Acute Myocardial Infarction." *Environmental Health Perspectives* 121 (2). National Institute of Environmental Health Science: 192.

Mosteller, Frederick, and John Wilder Tukey. 1977. "Data Analysis and Regression: A Second Course in Statistics." *Addison-Wesley Series in Behavioral Science: Quantitative Methods.*

Ord, Keith. 1975. "Estimation Methods for Models of Spatial Interaction." *Journal of the American Statistical Association* 70 (349). Taylor & Francis Group: 120–26.

Pace, R Kelley. 1997. "Performing Large Spatial Regressions and Autoregressions." *Economics Letters* 54 (3). Elsevier: 283–91.

Pelletier, Bruno, R Santer, and Jerome Vidot. 2007. "Retrieving of Particulate Matter from Optical Measurements: A Semiparametric Approach." *Journal of Geophysical Research: Atmospheres* 112 (D6). Wiley Online Library.

Racine, Jeff. 2000. "Consistent Cross-Validatory Model-Selection for Dependent Data: Hv-Block Cross-Validation." *Journal of Econometrics* 99 (1). Elsevier: 39–61.

Roberts, David R, Volker Bahn, Simone Ciuti, Mark S Boyce, Jane Elith, Gurutzeta Guillera-Arroita, Severin Hauenstein, et al. 2017. "Cross-Validation Strategies for Data with Temporal, Spatial, Hierarchical, or Phylogenetic Structure." *Ecography.* Wiley Online Library.

Robinson, George K. 1991. "That Blup Is a Good Thing: The Estimation of Random Effects." *Statistical Science.* JSTOR, 15–32.

Rosenblatt, Jonathan D. 2017. "R (Bgu Course)."

Rosenblatt, Jonathan D, and Boaz Nadler. 2016. "On the Optimality of Averaging in Distributed Statistical Learning." *Information and Inference: A Journal of the IMA* 5 (4). Oxford University Press: 379–404.

Schmidberger, Markus, Martin Morgan, Dirk Eddelbuettel, Hao Yu, Luke Tierney, and Ulrich Mansmann. 2009. "State-of-the-Art in Parallel Computing with R." *Journal of Statistical*

Software 47 (1).

Schott, James R. 2016. *Matrix Analysis for Statistics.* John Wiley & Sons.

Shao, Jun. 1993. "Linear Model Selection by Cross-Validation." *Journal of the American Statistical Association* 88 (422). Taylor & Francis: 486–94.

Stein, Michael L. 2012. *Interpolation of Spatial Data: Some Theory for Kriging.* Springer Science & Business Media.

Telford, RJ, and HJB Birks. 2009. "Evaluation of Transfer Functions in Spatially Structured Environments." *Quaternary Science Reviews* 28 (13). Elsevier: 1309–16.

Tobler, Waldo R. 1970. "A Computer Movie Simulating Urban Growth in the Detroit Region." *Economic Geography* 46 (sup1). Taylor & Francis: 234–40.

Trachsel, Mathias, and Richard J Telford. 2016. "Estimating Unbiased Transfer-Function Performances in Spatially Structured Environments." *Climate of the Past* 12 (5). Copernicus GmbH: 1215–23.

Vitter, Jeffrey Scott. 2001. "External Memory Algorithms and Data Structures: Dealing with Massive Data." *ACM Computing Surveys (CsUR)* 33 (2). ACM: 209–71.

Wang, Chun, Ming-Hui Chen, Elizabeth Schifano, Jing Wu, and Jun Yan. 2015. "Statistical Methods and Computing for Big Data." *arXiv Preprint arXiv:1502.07989.*

Wang, Yan, Itai Kloog, Brent A Coull, Anna Kosheleva, Antonella Zanobetti, and Joel D Schwartz. 2016. "Estimating Causal Effects of Long-Term Pm2. 5 Exposure on Mortality in New Jersey." *Environmental Health Perspectives* 124 (8). National Institute of Environmental Health Science: 1182.

Weiss, Robert E. 2005. *Modeling Longitudinal Data.* Springer Science & Business Media.

Yanosky, Jeff D, Christopher J Paciorek, Joel Schwartz, Francine Laden, Robin Puett, and Helen H Suh. 2008. "Spatio-Temporal Modeling of Chronic Pm 10 Exposure for the Nurses' Health Study." *Atmospheric Environment* 42 (18). Elsevier: 4047–62.

Zeger, Scott L, Duncan Thomas, Francesca Dominici, Jonathan M Samet, Joel Schwartz, Douglas Dockery, and Aaron Cohen. 2000. "Exposure Measurement Error in Time-Series Studies of Air Pollution: Concepts and Consequences." *Environmental Health Perspectives* 108 (5). National Institute of Environmental Health Science: 419.