

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 | 3 |
| 2 | Abstract | 3 |
| 3 | Introduction | 3 |
| 4 | Study Approach | 5 |
| 4.1 | Challenges | 5 |
| 4.2 | Research Objectives | 6 |
| 5 | Research Framework | 6 |
| 5.1 | Performance Estimation | 6 |
| 5.1.1 | The Loss Function | 7 |
| 5.1.2 | The Folding | 9 |
| 5.2 | Statistical Model | 10 |
| 5.2.1 | Mixed Models | 12 |
| 5.2.2 | GLS | 13 |
| 5.2.3 | The Errors Variance-covariance Matrix | 14 |
| 5.2.4 | Estimation Perspective | 18 |
| 5.2.5 | Regularization | 20 |
| 5.3 | Example: Validation and Estimation in Epidemiological Perspective | 21 |
| 5.4 | Computational Challenges | 21 |
| 5.4.1 | Sparse Representations | 22 |
| 5.4.2 | Memory Efficiency | 22 |
| 5.4.3 | Parallel Computing | 22 |
| 6 | Preliminary Results | 22 |
| 7 | Appendix | 23 |
| | References | 24 |

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. These assessments are then used as covariates in epidemiological studies; hence, their reliability is of high importance.

However, some complex issues arise when analyzed data have 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: (1) Propose an extension to classical methods that estimate prediction error (such as Cross Validation) for exposure assessment models with spatio-temporal structured data. (2) Improve models prediction performances while reducing its complexity using linear models. (3) Cope with the 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 field of environmental exposure and geo-statistics; not only in 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 exposure 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 common indicators for pollution levels [q][1]. However, PM mass concentrations are measured at ground monitoring stations, therefore highly limited in terms of spatial coverage. In order to provide PM assessments in geographical areas where no measurements are available some statistical methodology is needed.

PM levels show high spatio-temporal variation (Pelletier, Santer, and Vidot 2007); therefore, geostatistical interpolation methods in which spatial PM levels are modeled as opposed to some smoothness minimizer polynomial spline may introduce 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 to train models to predict pollution levels at any location ([q][14],[q][15]). The LUR methods provided more accurate exposure assessments at unmonitored locations; However they were significantly limited mostly since the included covariates are generally not time varying, 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 rising and new approaches were needed.

The increasing in availability of satellite data has enabled environmental scientists to harness remote sensing technologies to PM concentration assessments. Satellite-based *aerosol optical depth* (AOD) 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. AOD retrievals were found to be associated with ground PM measurements in different geographical areas. The use of AOD as a covariate in studies of PM concentration estimation has increased substantially over the past few years (e.g., (Kloog et al. 2014); (Chudnovsky et al. 2014); [ref]).

In recent studies, ground measured PM data were integrated with repeated AOD measurements 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 ([ref]; [ref]). 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 (see for instance [itai1][PM,AOD,MixedModel]; [itai2]). 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 ([q][good performance]; [q][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 described above procedure found statistically significant correlation between PM and variety of health related measurements such as adverse birth outcomes (Kloog et al. 2012); natural-cause mortality (Y. Wang et al. 2016); Acute Myocardial Infarction Madrigano et al. (2013)] and more.

In the following we introduce some of the challenges in satellite-based PM assessments models

(4.1), and the research objectives associated with them (4.2). In section 5.1 we provide a detailed review of the validation methods appropriate for the data with dependent structure that can be developed and fit. In section 5.2 we discuss the model statistical framework and present a generalized modification. In section 5.4 we show 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

4.1

The mission of PM exposure levels assesment is actually not well defined without additional setting. 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 analyses of these datasets requires a proper modeling of the dependence between observations. There are several approaches to model the dependency structure, each with its own pros and cons. Obviously, any selected approach has its own influence on the epidemiological results.

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, so the estimates can be expressed in matrix notation terms. Generally these term are easier to compute, depending on the structure of the data. Also, different estimation approaches could affect not only the computation time, but may also lead to significantly different results.

something about sparse representation, parallel computing and memory efficiency

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 generalized least squares (GLS) model and suggest several methods for estimating its required covariance matrix while modeling the spatio-temporal dependencies. Moreover, a regularization of the GLS model would also be investigated.
- Present an improvement in computation memory use, taking advantage of the fact that linear models consume less computational resources. Also, we would employ some of the recent achievements from the fields of sparse representation, parallel computing and memory efficiency in our framework.

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.

Furthermore, 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, usually in some form of the *sum of squares*. 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 assess the effect of air pollution on health, 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 closest to the truth.

It seems that when one is willing to perform a CV procedure in this settings, he faces two different hurdles: (1) choosing the right loss function so that it meets the objectives of the study, and (2) avoiding CV overfitting due to spatio-temporal dependency by choosing the folding scheme. We will break down this challenge into two decisions that must be made in accordance with the defined goals and the data structure.

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

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

$$\mathcal{L}_{uw} = (y - f(x))' (y - f(x)), \quad (1)$$

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

$$\mathcal{L}_w = (y - f(x))' \mathbf{W} (y - f(x)). \quad (2)$$

One option is to set the weights matrix \mathbf{W} as the inverse of the data variance-covariance 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 they are actually equivalent to only one observation.

Conversely, in 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 accurate 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 .

A typical epidemiologic study would compare some health indices in places experiencing different pollution levels. For instance, assume that the epidemiologist use the average values of the predicted PM levels \hat{y} in the city and in the village as explanatory variables. Also assume that the response variable is the morbidity rate in each of the locations.

Note that the experimental unit in the epidemiological regression is a geographical 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 accuracy in the village to get high accuracy in the city, since that both locations have the same weight in the epidemiological regression. Instead, he prefer that the data would be accurate 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, note that 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 argue that the loss function which evaluates model performance should be 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. Independency 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); [ref[]]. In other words, the estimated errors are downward biased, 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 version 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 version of LOO, *h-block CV* has proven to be asymptotically inconsistent (see 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 observations 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

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

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

Apparently CV approaches appropriate for spatially dependent data receive 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 that the scheme of h-block can be adopted for the spatial case by omitting observations within a radius of h of the test set (also referred as spatial-LOO). 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 particularly 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 may 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 the choice of block length in an iterative fashion under suitable assumptions.

5.2 Statistical Model

Although dependency among observations is a common phenomenon in observational studies, it violates one of the standard statistical assumptions, and challenging many classical statistical techniques. Geographical and environmental data generally show both spatial and temporal dependency between observations, known as autocorrelation. Standard regression techniques ignore this dependency structure and usually lead to unstable estimates [ref], bad predictions (N. A. Cressie 1993) (p.435), and biased results [ref] resulting in erroneous conclusions.

In the last decades there has been an explosion of research in dependence data modeling. Logitudinal modeling is perhaps the most familiar one. It mostly focuses on the dependence among observations over time such as time series, and is commonly used in many fields such as biology, economics and more. A number of approaches to modeling longitudinal data have been proposed in the statistical literature. A very common approach is to use a parameterized covariance model. parameterized models assume an underlying structure of stationary stochastic process, which can be described 'using a with small number of

parameters. such as structures are *autoregressive* (AR), or *moving average* (MA) models. For a comprehensive review see 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 geography (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) introduce 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 describe the spatial arrangement of the geographical units and can be assumed or estimated. in his seminal book, Anselin (2013) 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 referred it as a special case of the more general *spatial Durbin model* (SDM). Another method that have recently shown to be effective is *Bayesian hierarchical* modeling approaches (see Banerjee, Carlin, and Gelfand (2014)). These models handle complex relationships such as multi-level data by estimating the parameters of the posterior distribution using the bayesian methods.

Spatio-temporal data enable the researcher to take advantage of time-space 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 need to be accounted. It is an emerging research 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 the estimated variance-covariance matrix of the error terms to efficiently estimate model's parameters in the presence of dependency. Since it is a linear model, the GLS has some nice properties as we will discuss later. Although 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 effect 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 should be modeled?". We state that, as far as regressions are concerned, 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 argue 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, in choosing the statistical model we should also consider its goal. The purpose of the model in our case is predication rather than inference. For this reason, we might consider reducing prediction error in cost of biased model, using regularization. Suitable regularization approaches for regressions are Shrinkage methods such as *Ridge regression* or *Lasso*. It should be noted that while such procedures are relatively easy to adapt for linear models, It might be more complex for mixed effect models. [ref]

5.2.1 Mixed Models

Mixed models, sometimes referred as Hierarchical models, are a class of statistical models suited for the analysis of structured data. Mixed models are particularly useful when observing 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 introducing multilevel random effects. In these 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 areas level, they might be spatial grid cells. The model assumes that observations between clusters are independent, while observations 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 measurements 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 measurements across different geographic locations. This cluster-specific uniqueness is reflected in an estimated posteriori coefficient and referred 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 assessment literature is the *linear mixed effect* (LME) model that was originally developed by Laird and Ware (1982). It can be formalized as:

$$y_j = X_j\beta + Z_jb_j + \varepsilon_j \quad j = 1, \dots, T \quad (4)$$

where:

j represent a cluster, s_j is the number of observations in cluster j , and T is the number of clusters; y_j is an $s_j \times 1$ vector of responses of the j th 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_i is an $s_j \times 1$ independent and identically distributed (*iid*) error terms vector. each component of ε_j (obsevation in the j th cluster) is assumed to have mean zero and variance σ^2 .

The compressed form of N equations is:

$$y = X\beta + Zb + \varepsilon \quad (5)$$

or,

$$y = X\beta + \eta \quad (6)$$

where

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

The model assumes that $E(\eta) = 0$. Note that $\text{Var}(\eta)$ is an $N \times N$ covariance matrix, where $N = \sum_{j=1}^T s_j$. Let us define $V = \text{Var}(\eta) = E(\eta\eta')$. V has the following block diagonal form:

$$V = \sigma^2 \begin{bmatrix} I_{s_1} + Z_1 D Z_1' & 0 & 0 & 0 \\ 0 & I_{s_2} + Z_2 D Z_2' & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & I_{s_T} + Z_T D Z_T' \end{bmatrix} \quad (8)$$

where I_{s_i} is the identity matrix in s_i dimension.

In other words, without further assumptions (i.e residuals autocorrelation), the LME can be considered as the familiar linear model, exept that it assume that the residuals covariance matrix V follows a specific structure. In particular, V is forced 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 the following matrix notation:

$$y = X\beta + \varepsilon. \quad (9)$$

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

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

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

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

Clearly, the quality of the GLS estimator 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 GLS includes as its special cases various specific models that are determined by the estimated error terms covariance matrix.

In addition, we want to emphasize that GLS estimator is also the minimizers of the squared mahalanobis length of the residual vector $y - X\beta$. This is particularly important since we might want to choose an estimator that minimizes the *weighted loss function* which is described above as an option to measure model performance using the estimated Σ :

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

Note that GLS 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 estimated Σ , strongly dependent observations, which usually have highly correlated errors, will have less impact on the values of the estimators 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 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 function of $q \in \{1, \dots, N(N+1)/2\}$ parameters, where N is the number of error terms. As an initial step our concentration will be in *stationary* covariance functions $f : \mathbb{R}^d \times \mathbb{R} \rightarrow \mathbb{R}$, defined by N. Cressie and Wikle (2015) as:

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

We define the covariance model as $\Sigma_C(\theta)$, where C refers to the chosen covariance function and θ to the distinct $q \times 1$ parameters vector that need to 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 by the model (i.e. AR, SEM, etc.), but parsimonious, meaning that 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-scaled identity matrix, sometimes called *spherical error variance* matrix (Hayashi 2000). This structure assumes *homoscedasticity* and no autocorrelation, and requires estimation of only one parameter. $\Sigma_u(\theta)$ 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 discuss some procedures of temporal and spatial modeling separately and further examine a space-time integrated model. We will stick to a spatio-temporal framework in which $i \in 1, \dots, S$ indicates a spatial unit and $j \in 1, \dots, T$ indicates a time unit.

5.2.3.1 Temporal perspective

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

Consider the model in (9), only with the extension:

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

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

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

We ignore here the spatial pattern and assume that the process is spatially constant. That is, at every spatial unit i the $T \times T$ covariance matrix is the same:

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

where I_S is the $S \times S$ identity matrix and $\tau^2 = \text{Var}(\varepsilon_{ij}) = \frac{\sigma_\varepsilon^2}{1-\rho^2}$. Note that this AR(1) covariance model requires the estimation of 2 parameters, and in general $p + 1$ parameters 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)). It is defined as following:

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

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 the *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 ε :

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

where W_{ik} is the $i - k$ th element in the weight matrix W , and ν_i is an iid white noise with $E(\nu_i) = 0$ and $\text{Var}(\nu_i) = \sigma_\nu^2$. Note that we ignore here temporal effects, assuming the same structure for each time unit j . $\sum_{k=1}^S w_{ik} \varepsilon_k$ is called the *spatial lag*, since it represent a linear

combination of values of the error terms ε constructed from regions (spatial units) that neighbor ε_i . λ is the correlation between the errors and their spatial lag. In matrix notation the $S \times 1$ vector ε can be represented as:

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

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

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

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

- *k-nearest neighbors*:

$$W_{ik} = \begin{cases} 1 & , u \in N_k(v) \\ 0 & , otherwise \end{cases} \quad (21)$$

- *Radial Distance*:

$$W_{ik} = \begin{cases} 1 & , 0 \leq d_{ik} \leq L \\ 0 & , otherwise \end{cases} \quad (22)$$

- *Power Distance*:

$$W_{ik} = \frac{1}{d_{ik}^\alpha} \quad (23)$$

Another approach to model the covariance is to specify 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 specify the full covariance matrix. Some instances for the K_{ik} function are:

- *Negative exponential*:

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

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

- *Spherical*:

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

The *variogram* (see N. A. Cressie (1993)) is another approach that can be used to describe 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 field values (here values of errors) at spatial locations s and $s + h$:

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

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 easily 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 $t, r \in \mathbb{R}$

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

where $C^{(s)}$ and $C^{(t)}$ are the spatial and temporal covariance function. Under spatio-temporal separability, the covariance matrix can be written as a *kroncker product* of the separately 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, a fundamental issue in building an errors covariance matrix, 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)). Note that in this case the error covariance matrix is quite sensitive to the OLS regression, hence it is important that the functional structure of this regression will result in residuals which reflect the true error.

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 distribution of the error, specifically:

$$\varepsilon_i \sim \mathcal{N}(0, \sigma^2 I) \quad b_i \sim \mathcal{N}(0, \sigma^2 D) \quad (28)$$

and $I = I_{s_j}$.

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

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

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

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

This log likelihood function involves matrix inverse and determinante which might be a difficult task if the matrix dimension is high. However, some dimension reduction formulation can be employed in order to make calculation easier, se 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:

$$\Sigma^{-1} = P P', \quad (31)$$

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

Note that mltiplying both sides of (9) by P yields:

$$\tilde{y} = \tilde{X} \beta + \tilde{\varepsilon}, \quad (32)$$

where, $\tilde{y} = P y$, $\tilde{X} = P X$ 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* (GLSE) β_{GLS} is achieved using (32) and the OLS:

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

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 common with respect to GLS (Fomby, Hill, and Johnson 2012). The FGLS estimation proceeds in two stages which can be repeated for several iterations: (1) the model is estimated assuming Σ is known, and the residuals are then used to construct the error covariance matrix by estimating θ . (2) 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].

Assuming normal distribution of the error term ε , the log-likelihood function of (9) can be written as:

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

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 distribution 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

Although 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*). 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 methods is that they usually involve complex numerical optimization, therefore are computational intensive in relation to classical regularization methods

5.2.5.2 Regularization in GLS

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

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

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:

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

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

5.3 Example: Validation and Estimation in Epidemiological Perspective

EIV.Rmd file goes here

ask johnathan if necessary

=====

TODO:

5.4 Computational Challenges

- why gls is easy to compute and to parallelize
- covariance matrix: the computational burden increases with sample size. small number of nonzero elements - reduce the computational burden considerably.

5.4.1 Sparse Representations

The covariance matrix: the main barrier to speedy computation of the estimates lies in the N by N nature of the spatial weighting matrix. Pace (1997) has suggested the use of sparse matrix techniques to facilitate the use of large samples. If Σ is specified so that the number of nonzero elements is relatively small, these methods can reduce the computational burden considerably.

5.4.2 Memory Efficiency

5.4.3 Parallel Computing

Yan et al. (2007) - Parallelizing MCMC for Bayesian spatiotemporal geostatistical models -

=====

6 Preliminary Results

7 Appendix

References

- 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.
- Anselin, Luc. 2013. *Spatial Econometrics: Methods and Models*. Vol. 4. Springer Science & Business Media.
- 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.
- 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.
- Burman, Prabir, Edmond Chow, and Deborah Nolan. 1994. "A Cross-Validatory Method for Dependent Data." *Biometrika* 81 (2). Oxford University Press: 351–58.
- 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.
- Elhorst, J Paul. n.d. *Spatial Econometrics: From Cross-Sectional Data to Spatial Panels*. Springer.
- 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.
- Genton, Marc G. 2007. “Separable Approximations of Space-Time Covariance Matrices.” *Environmetrics* 18 (7). Wiley Online Library: 681–95.
- 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.
- Kariya, Takeaki, and Hiroshi Kurata. 2004. *Generalized Least Squares*. John Wiley & Sons.
- Kloog, Itai, Alexandra A Chudnovsky, Allan C Just, Francesco Nordio, Petros Koutrakis, Brent A Coull, Alexei Lyapustin, Yujie Wang, and Joel Schwartz. 2014. “A New Hybrid Spatio-Temporal Model for Estimating Daily Multi-Year Pm 2.5 Concentrations Across North-eastern Usa Using High Resolution Aerosol Optical Depth Data.” *Atmospheric Environment* 95. Elsevier: 581–90.
- 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.
- 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.
- 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 Pm_{2.5} 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-Aroita, 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.

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.

Wang, Yan, Itai Kloog, Brent A Coull, Anna Kosheleva, Antonella Zanobetti, and Joel D Schwartz. 2016. “Estimating Causal Effects of Long-Term Pm_{2.5} 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.

Yan, Jun, Mary Kathryn Cowles, Shaowen Wang, and Marc P Armstrong. 2007. “Parallelizing Mcmc for Bayesian Spatiotemporal Geostatistical Models.” *Statistics and Computing* 17 (4). Springer: 323–35.

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.