Predicting French Air Pollution

— A spatio-temporal analysis of particulate matter over Metropolitan France

James Keane

Advisor: Peter McCullagh

Approved _____

Date _____

February 21, 2022

Abstract

The forces that generate environmental phenomena are many and complex. Common geo-statistical approaches to predicting such phenomena involve regressing on related maps of land-use regressors, and interpolating observed samples across space and time. In this paper, we review the method of unbiased, linear interpolation known as Kriging, and apply it to a dataset of French $PM_{2.5}$ readings in 2020. We harness the spatio-temporal dependencies of the pollution process with ordinary and residual Kriging methods to interpolate readings across the whole of Metropolitan France.

Contents

1 Introduction **4**

2 Data **5**
2.1 Study Area . . . 5
2.2 Pollution Readings . . . 5
2.3 Land-use Regressors . . . 6
2.3.1 Road Covariates . . . 6
2.3.2 Land Use/Cover Proportions . . . 7
2.3.3 Meteorological Data . . . 7
2.3.4 Population Density . . . 7
2.3.5 Elevation . . . 8
2.4 Satellite Readings . . . 8
2.4.1 Aerosol Optical Depth . . . 8
2.4.2 Normalized Difference Vegetation Index . . . 9
2.5 Data Pre-processing . . . 10
2.5.1 Projection . . . 10
2.5.2 Matching Spatial Resolution . . . 10
2.6 Full Variable List . . . 11

3 Methodology **12**
3.1 Assumptions . . . 12
3.1.1 Stationarity . . . 12
3.1.2 Isotropy . . . 12
3.2 Variograms . . . 13
3.3 Covariance Structures . . . 14
3.3.1 Separable Models . . . 15
3.3.2 Product-sum Models . . . 15
3.3.3 Metric Models . . . 15
3.3.4 Sum-metric Models . . . 15
3.3.5 Simple Sum-metric Models . . . 16
3.4 Kriging Types . . . 16
3.4.1 Simple Kriging . . . 16
3.4.2 Ordinary Kriging – Unknown, Constant Mean . . . 17
3.4.3 Kriging with Drift – Unknown, Varying Mean . . . 17
3.4.4 Residual Kriging – Separate Drift Estimation and Residual Interpolation 18
3.5 Kriging with Large Datasets . . . 19
3.5.1 Model Fitting Through the Semivariogram . . . 19
3.5.2 Local Neighborhood Kriging . . . 20
3.6 Model Evaluation . . . 20

4 Modeling Process **22**
4.1 Holdout Set Creation . . . 22
4.2 Land-use Regression Structure . . . 22
4.2.1 Hierarchical Lasso . . . 22
4.2.2 Fitted Model Summary . . . 24
4.2.3 De-trended Residuals . . . 25
4.3 Spatio-Temporal Dependence . . . 26
4.3.1 Checking for Anisotropy . . . 26
4.3.2 Semivariogram Modeling . . . 26
4.4 Model Selection . . . 27

5 Results **29**
5.1 Hold-out Set Predictions . . . 29
5.2 Full Grid Predictions . . . 31
5.2.1 Ordinary and Residual Kriging Comparison . . . 31
5.2.2 Kriging Standard Error Back-Transformation Changes . . . 32

6 Conclusions **33**

A Appendix **34**
A.1 Ordinary Kriging Model . . . 34
A.2 Formulas and Derivations . . . 35
A.2.1 LCC Coordinate Formulas . . . 35
A.2.2 Semi-variogram Relation to Covariance . . . 36

B Additional Visualizations **36**
B.1 Log $PM_{2.5}$ Holdout Error Plots . . . 36
B.2 Study Area Log $PM_{2.5}$ Predictions . . . 38

1 Introduction

Air pollution provides a major concern to the modern world with long-term exposure to hazardous aerial substances posing a threat to both climate and population health. Of pollutant classes, fine particulate matter under 2.5 microns in width – commonly labeled $PM_{2.5}$ – poses one of the greatest risks to human health, with around 3.2 million deaths per year attributed to the particles ([21]). Continued exposure to high levels of $PM_{2.5}$ is known to exacerbate pre-existing breathing conditions, while also being an aggravator to the elderly and to those with heart conditions.

In this paper, we consider the levels of $PM_{2.5}$ in France. While the nation is far from the worst sufferer of pollution, issues with $PM_{2.5}$ placed them in the top three European countries for deaths that could be directly attributed to the pollutant in 2017. ([7]). Further years have seen the European Commission take the nation to court over their excessive pollution rates ([8]). With these clear ramifications, it becomes imperative to have readings for air quality over time and space to better track pollution trends. This paper focuses on creating a statistical model capable of providing estimates of ambient air quality concentrations of $PM_{2.5}$ across the nation for the year 2020.

Previous approaches to modeling air pollution include spatial smoothing techniques such as Kriging, as well as more standard regression models typically labeled LURs or "land use regressions." ([9], [10]) The former uses observed pollution readings to interpolate to unobserved locations while the latter harnesses other environmental characteristics and GIS variables as a more indirect method of estimating pollution.

In this paper, we consider combining the explanatory ability of commonly available land user regressors with spatio-temporal Kriging to create accurate, daily estimates of $PM_{2.5}$ at unobserved locations in France. Section 2 covers the available covariates with possible predictive ability, while sections 3 and 4 provide background on the Kriging process, and the model-building process, respectively. While the forces that affect ambient air pollution are many and complex, in section 5 we find that the marriage of land-use regression and Kriging interpolation is able to out-perform either individual component in predicting $PM_{2.5}$.

2 Data

2.1 Study Area

Our area of interest is commonly labeled as *Metropolitan France*. This includes the twelve regions of the country on continental Europe, as well as the island region of Corsica. We divide the area of 543,940 km^2 into even cells of roughly 5x5km each.

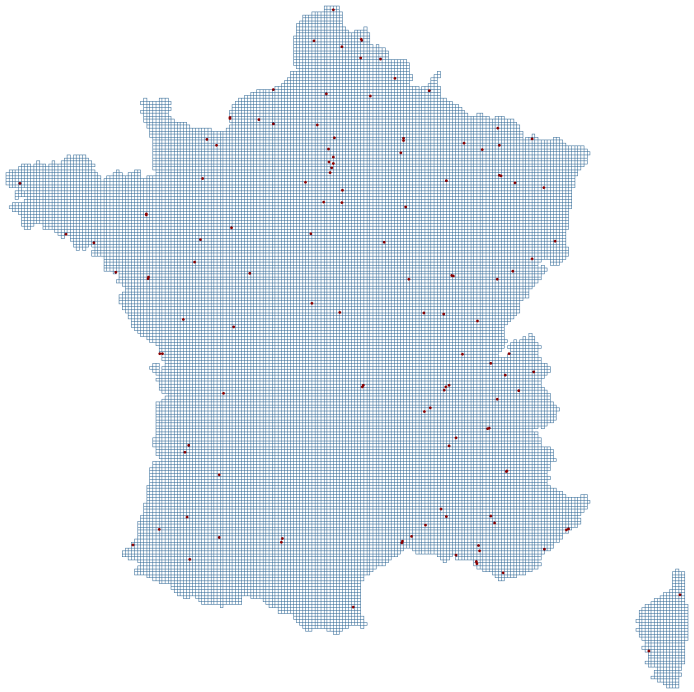


Figure 1: Metropolitan France, divided into 5x5km grid cells. $PM_{2.5}$ sensor locations marked in red.

2.2 Pollution Readings

Sensor readings of $PM_{2.5}$ are available from OpenAQ, an open-source air quality data aggregator. All available sensors covering this metric of air pollution within the study area are supported by the French branch of the European Environment Agency (EEA). We download averaged daily $PM_{2.5}$readings for 134 available sensors. Of these, 8 sensors are not considered in the modeling process due to a substantial fraction of missing values. The remaining 126 sensors – with locations marked in figure 1 – have a daily completion rate of at least 85% across 2020.

The upper target value for mean $PM_{2.5}$ concentration within a 24-hour period, as designated by the World Health Organization, is 25 $\mu g/m^3$ ([11]). Other standards created by the U.S. Environmental Protection Agency (EPA) consider exposures above 35 $\mu g/m^3$ as unhealthy to sensitive groups, and above 55 $\mu g/m^3$ as unhealthy to all ([12]).

French particulate matter readings in 2020 fell largely within the healthy limits between 0 and 20 $\mu g/m^3$. With the distribution being purely non-negative and right-skewed by more extreme pollution events. A log-transformation (figure 2) provides a more normal shape to the observations, and is used in the modeling process.

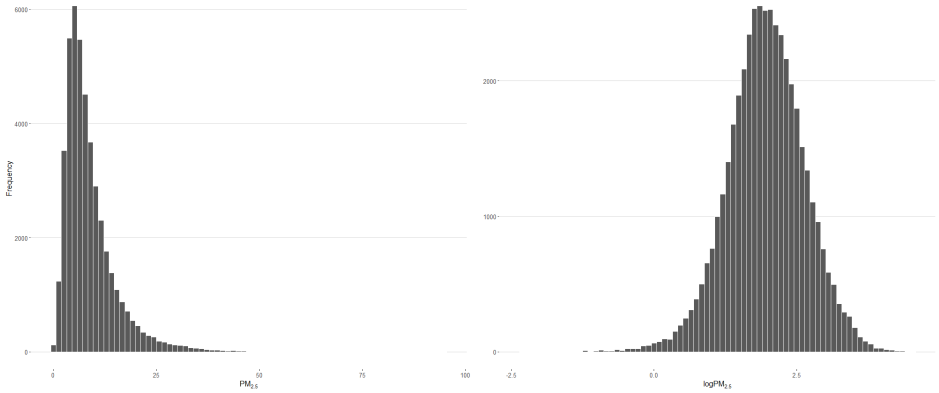


Figure 2: 2020 $PM_{2.5}$ Readings pre(left) and post(right) log transformation

2.3 Land-use Regressors

2.3.1 Road Covariates

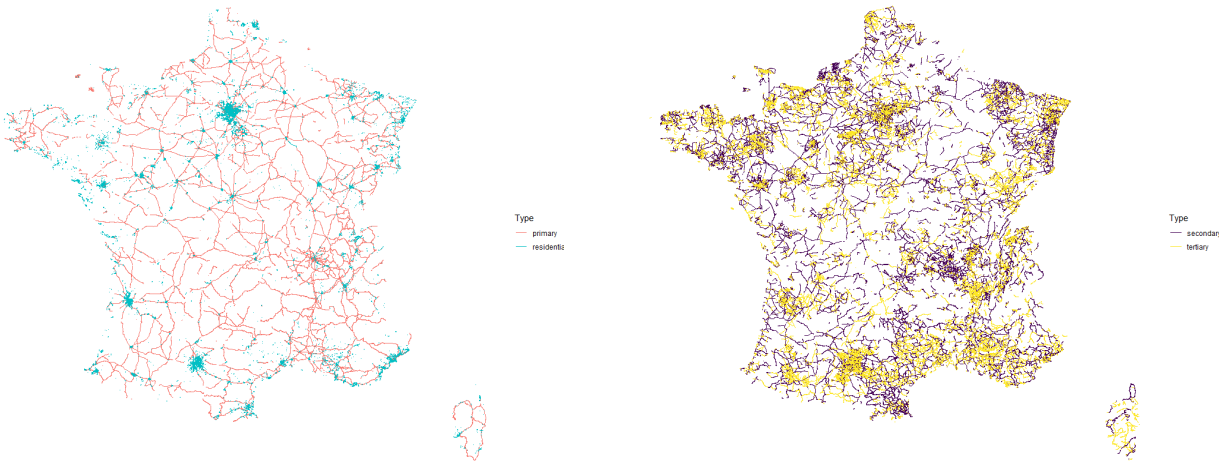


Figure 3: Maps of France by Road Type

Vehicular transmissions have been shown to be strong contributors to ambient $PM_{2.5}$ production ([6]). In lieu of granular traffic data, we use various spatial relationships with roads to indirectly include traffic effects into our modeling. Using road data from OpenStreetMap, an open source geographic database, we utilize measures of the following road types:

- Major Roads
  - Motorways linking larger towns and cities
- Secondary Roads
  - Highways linking towns
- Tertiary Roads
  - Roads linking smaller towns and villages
- Residential Roads
  - Roads with direct access to housing, often lined with houses.

For each of the four road types (3), we measure the proximity of the nearest road to each pollution sensor. Additionally, we measure the density of roads within each 5km grid cell. To ensure remote locations on the map (without roads) are closer to our sample distribution, we take the square root of all our road measures.

2.3.2 Land Use/Cover Proportions

Land cover categories are available through the Environmental Systems Research Institute (Esri) at a 10x10m cell resolution. Cells of water, trees, crops, vegetation, and built area were aggregated as proportions within each of the study area's designated 5x5km grid cells. Other Esri land use cell types of snow, shrub, clouds, and flooded vegetation are not considered due to their sparsity in France.

2.3.3 Meteorological Data

Previous analyses have found weather data to be quite predictive of $PM_{2.5}$ concentrations ([13], [14]). To incorporate weather data in our modeling, we include measures for average temperature, relative humidity, average wind speed, average sea-level pressure, and total precipitation provided at a 0.1x0.1° resolution by the European Climate Assessment & Dataset Project (ECAD).

2.3.4 Population Density

French population data for 2020 is included from WorldPop, an open access archive of global population datasets. The density is at a resolution of 1x1km, and is mean-aggregated to the study area grid cell size of 5x5km.

2.3.5 Elevation

Elevation values for the $PM_{2.5}$ sensors and the study area grid cells are accessed from EU-DEM, a digital surface model that covers Europe at a 25 meter resolution. The majority of France falls below one thousand meters above sea-level, with more extreme elevations coming in the Alps in the south east of the nation. As the sensor elevations do not surpass one thousand meters, we log-transform the elevation covariate in our modeling to make the distribution of our observed sampling locations and the rest of the study area more similar.

2.4 Satellite Readings

2.4.1 Aerosol Optical Depth

Optical depth, or optical thickness, is a measure of tiny particles suspended in the air between Earth and the top of the atmosphere. Effectively, higher values of optical depth indicate more gases or particles in the air scattering light. Aerosol optical depth (AOD) is the portion of that value that is caused by particulate matter buildup, and therefore has a natural positive association with ground-level $PM_{2.5}$readings. We use NASA product MCD19A2 to include daily AOD readings at wavelength 470nm at 1x1km spatial resolution.

Due largely to enduring cloud cover obfuscating the satellite observations, optical depth can often not be calculated for a given location at a given day. In our dataset, a total of 43% of grid cell location and day pairs are missing observed AOD readings. Other analyses have found that maintaining the measure in predictive modeling of $PM_{2.5}$has been useful, in spite of needing to mass-impute the missing values ([16], [18]).

Gaps in our AOD readings are filled by values from the Copernicus Atmosphere Monitoring Service reanalysis (CAMS). This project provides estimates for AOD at a 470nm wavelength twice a day and at a spatial resolution of 10x10km. Our imputation then averages the twice daily estimates, and approximates the measure for the higher resolution study area grid cells by bilinear interpolation. An example of the imputation is visualized in Figure 4.

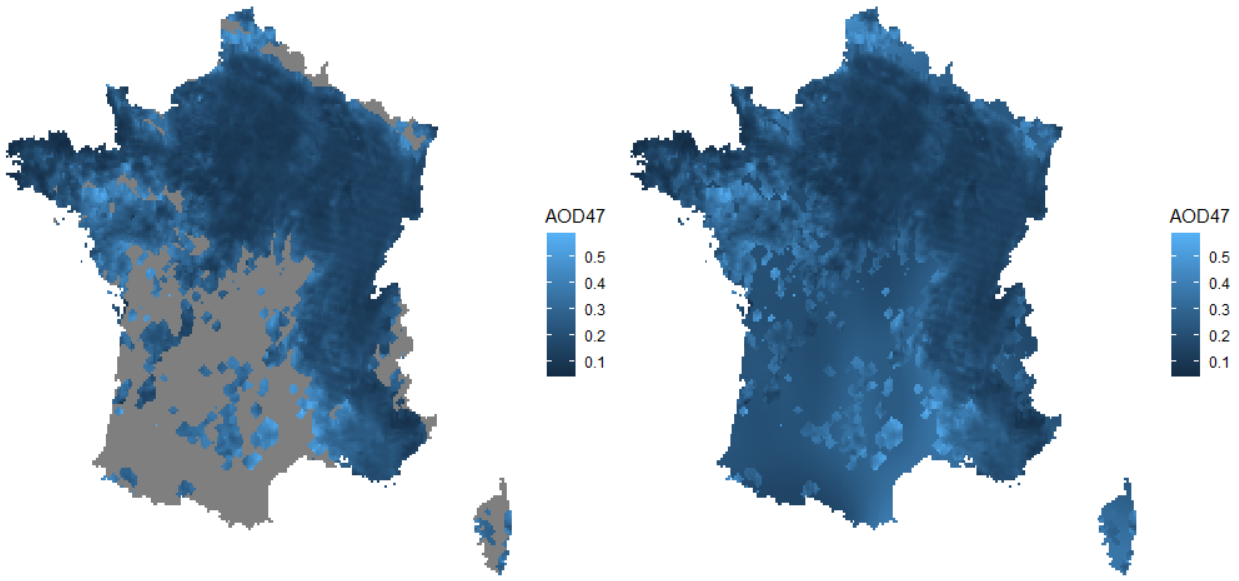


Figure 4: AOD measures for 04/16/2020 pre(left) and post(right) imputation

2.4.2 Normalized Difference Vegetation Index

NDVI (Normalized Difference Vegetation Index) is commonly considered a quantification of the level of vegetation in an area. Using satellite imagery, NDVI is computed as a ratio of near infrared (NIR) and red light.

$$NDVI = \frac{NIR - Red}{NIR + Red}$$

This formulation forces NDVI to range from -1 to 1. More positive values indicate a lack of reflection, likely due to dense vegetation or greenery. Negative values indicate reflective surfaces, such as snow or water. Values around zero are typically considered to be urban settings, or settings with dead vegetation. This is reflected in Figure 5, as the average observed NDVI follows a sensible rise when vegetation is growing and fall when it is dying.

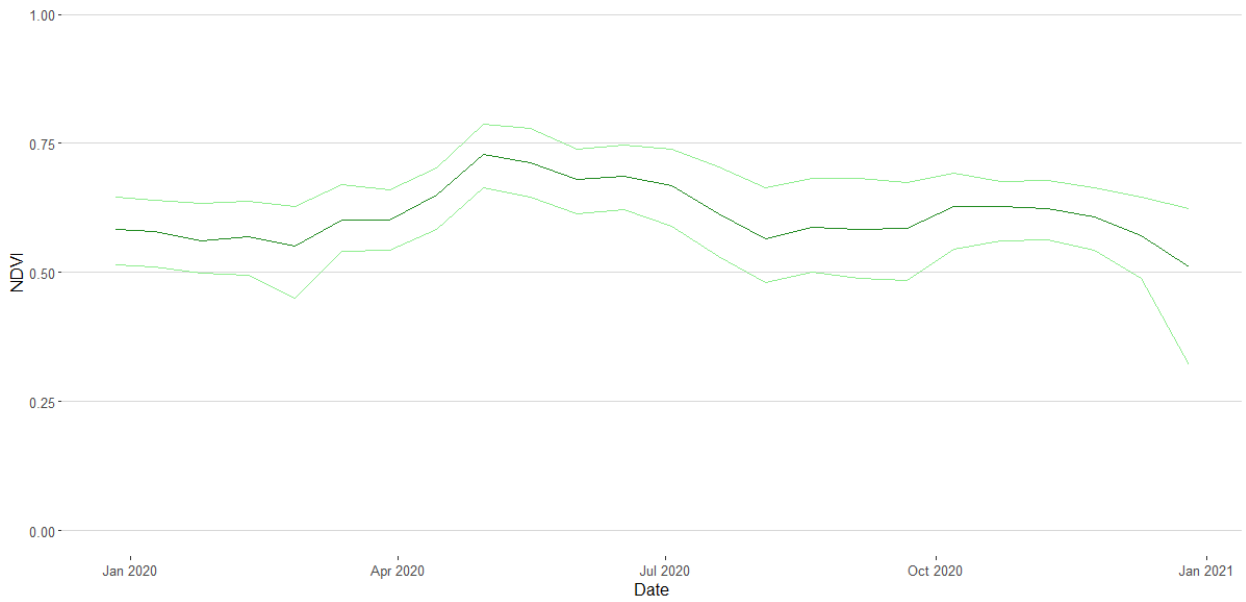


Figure 5: 2020 Time Series of Median, Q1, and Q3 NDVI in Metropolitan France

Similar to aerosol optical depth, satellite readings of NDVI are commonly obfuscated by cloud coverage. Since vegetation changes gradually, we include averaged NDVI across 16-day windows to include full coverage. NASA's MYD13A2 product supplies full coverage for 16-day NDVI at 1km resolution [1].

2.5 Data Pre-processing

2.5.1 Projection

Latitude and longitudes of all variables in the dataset are either accessed at or re-projected to Lambert conformal conic (LCC) 93 coordinates. This has the benefit of cartesian coordinates, allowing distance in meters to be computed directly with euclidean distance:

$$d((lon_1, lat_1), (lon_2, lat_2)) = \sqrt{(lon_1 - lon_2)^2 + (lat_1 - lat_2)^2)}$$

The translation formulas between standard latitude and longitude (WGS84) and the LCC-projected coordinates are available in Appendix section A.2.1.

2.5.2 Matching Spatial Resolution

Variables at higher spatial resolutions than the study area are matched to the 5x5km grid cells by simple aggregation or down-scaling. Aggregation functions used are mean and proportion aggregations based on the variable involved.

Upscaling covariates involves resampling lower-resolution data to match the 5x5km grid cells. This is done via bilinear interpolation, which uses a linear combination of the observations of the four nearest cell centers (denoted as (x_1, y_1), (x_1, y_2), (x_2, y_1), & (x_2, y_2)) to the location of interest $((x_0, y_0))$ to form an estimate.

$$f(x_0, y_0) = \frac{1}{(x_2 - x_1)(y_2 - y_1)}[x_2 - x_0, \quad x_0 - x_1] \begin{bmatrix} f(x_1, y_1) & f(x_1, y_2) \\ f(x_2, y_1) & f(x_2, y_2) \end{bmatrix} \begin{bmatrix} y_2 - y_0 \\ y_0 - y_1 \end{bmatrix}$$

2.6 Full Variable List

Table 1 summarises the full dataset of variables considered in the modeling process, and their initial resolutions.

Table 1: List of Variables and Definitions

Name	Units	Definition	Spatial Res.	Scaling
$PM_{2.5}$	$\mu g/m^3$	Ground-level Particulate Matter	Point	-
lonL	°	Longitude, LCC-projected	Point	-
latL	°	Latitude, LCC-projected	Point	-
DOY	-	Day of year	-	-
DOW	-	Day of week	-	-
Month	-	Month of year	-	-
Elevation	m	Altitude	25x25m	Mean Agg
Crop_prop	%	Land use, Crop proportion	10x10m	Prop Agg
Built_prop	%	Land use, Human structure proportion	10x10m	Prop Agg
Trees_prop	%	Land use, Forested proportion	10x10m	Prop Agg
Water_prop	%	Land use, Water proportion	10x10m	Prop Agg
Veg_prop	%	Land use, Low vegetation proportion	10x10m	Prop Agg
Pop_density	Pop/km	2020 Population Density	1x1km	Mean Agg
Hway_density	m	2020 Density of A1, A2, A3 Road types	5x5km	-
Res_density	m	2020 Density of Residential Roads	5x5km	-
Near_maj	m	Proximity to nearest A1 road	-	-
Near_sec	m	Proximity to nearest A2 road	-	-
Near_ter	m	Proximity to nearest A3 road	-	-
Near_res	m	Proximity to nearest residential road	-	-
AOD47	No unit	Aerosol Optical Depth at 470nm	1x1km	Mean Agg
NDVI	No unit	Normalized Differenced Vegetation Index	1x1km	Mean Agg
Humidity	%	Relative Humidity	0.1x0.1°	Bilinear
Temp	°C	Average temperature	0.1x0.1°	Bilinear
Pressure	hPa	Average sea-level pressure	0.1x0.1°	Bilinear
Precip	mm	Total precipitation	0.1x0.1°	Bilinear
Wind	m/s	Average wind speed	0.1x0.1°	Bilinear

3 Methodology

For the purposes of this analysis, we model the pollution readings of interest as observations of a Gaussian spatio-temporal field with a spatial domain that includes Metropolitan France, and a temporal domain that includes all days in 2020. This allows us to describe the mean and covariance of the process at any point in the domain, and to make predictions at unobserved spatio-temporal locations using Kriging interpolation of observed samples.

We formulate our process Z as an additive function of a deterministic mean (μ) and two independent, zero-mean Gaussian processes, ϵ and η. The former is an iid noise term covering the random noise in the pollution process. The latter process, η, covers spatio-temporal dependencies in the pollution process. Altogether, we formulate our process $Z(s,t)$, a function of spatial location s and time location t:

$$Z(s,t) = \mu(s,t) + \eta(s,t) + \epsilon(s,t)$$

3.1 Assumptions

3.1.1 Stationarity

Stationarity of a process requires that all moments of a distribution be constant across the entirety of the study domain. Various methods of Kriging assume *second-order*, or *weak stationarity*, which requires that the expected value of the process Z be the same at all time t and space s pairs in the domain D:

$$E[Z(s,t)] = \mu \quad \forall s,t \in D_{s,t}$$

Further, weak stationarity also requires that the covariance between any point pair is solely a function of the spatio-temporal lags h & τ and the direction of those lags, between the pair.

$$Cov(Z(s,t), Z(s+h, t+\tau)) = Cov(Z(0,0), Z(h,\tau)) = Cov(h,\tau)$$

For fields with non-constant expected values, the stationarity assumption of the mean can be relaxed by $de-trending$ the mean and assuming constant expected value residuals. This is a step performed in more complex Kriging processes.

3.1.2 Isotropy

A random field that is *isotropic* extends the weak stationarity property, requiring that the covariance between two points be solely a function of their directionless spatio-temporal distance.

$$Cov(Z(s,t), Z(s+h, t+\tau)) = Cov(||h||, |\tau|)$$

This uniformity in all directions makes isotropic covariances completely symmetric as the covariance for a spatio-temporal lag of twenty kilometers due west and one week in the

future is identical to the covariance for a lg of twenty kilometers due east and one week in the past.

The absence of isotropy is known as *anisotropy*.

Geometric anisotropy features the same covariance structure, but with different rates at which the covariance shrinks based on direction. This can be remedied by modifying coordinates to shrink or increase distances as needed to "unify" the covariance in all directions.

Zonal anisotropy is the instance in which the entire covariance form or parameters vary by direction, making it a more difficult problem to resolve ([20]).

Unaccounted for anisotropy has a notable effect on Kriging weights. As the degree of anisotropy increases, the disparity in Kriging weights assigned to two points equidistant to the prediction location should increase. An isotropic covariance structure will not account for this disparity, leading to worse predictions, and higher prediction variances.[[19]]

3.2 Variograms

Alongside the covariance matrix, dependencies across space and time are commonly modeled through the semi-variogram characterized through γ. This is defined as the variance of the difference of two points in the process domain:

$$2\gamma(s,t;s',t') = Var(Z(s,t) - Z(s',t'))$$

If the semivariogram of a process is solely a function of the spatio-temporal lag between two points, then a process is *intrinsically stationary* ([2]).

$$2\gamma(h,\tau) = Var(Z(s,t) - Z(s+h,t+\tau))$$

All second-order stationary processes with finite variance are intrinsically stationary, but the reverse does not hold. However, if there is second-order stationarity, the semivariogram can then be expressed directly as a function of the covariance:

$$\gamma(h,\tau) = C_z(0,0) - C_z(h,\tau)$$

With a provided covariance structure, the semivariogram can be defined by the following parameters:

- The *sill* is the γ value as the lag approaches ∞.
- The *range* is the lag at which the *sill* is reached. Point pairs at or beyond this lag are generally considered to be uncorrelated.
- At a lag of 0, the semivariance is naturally 0 as there is no difference between a point and itself. The *nugget* is the γ value as the lag approaches 0. A non-zero nugget indicates the amount of variability between two points that are almost at the same location on the spatio-temporal domain. This variability then cannot be explained by spatial or temporal correlations.
- The *partial sill* is the difference between the nugget effect and the full sill.

Two common models for the semivariogram (and the covariance) used in GIS settings are the (i) spherical:

$$C_{sph}(h) = \begin{cases} b(1 - \frac{3|h|}{2a} + \frac{|h|^3}{2*(a^3)}), & \text{for } 0 \leq |h| \leq a \\ 0 & \text{for } a < |h| \end{cases}$$

$$\gamma_{sph}(h) = \begin{cases} b(\frac{3|h|}{2a} - \frac{|h|^3}{2*(a^3)}), & \text{for } 0 \leq |h| \leq a \\ b & \text{for } a < |h| \end{cases}$$

and the (ii) exponential:

$$C_{exp}(h) = b * e^{-\frac{|h|}{a}} \quad a, b > 0$$

$$\gamma_{exp}(h) = b * (1 - e^{-\frac{|h|}{a}}) \quad a, b > 0$$

The former and latter lead to a quicker and more gradual marginal semivariogram growth respectively, as visible in Figure 6.

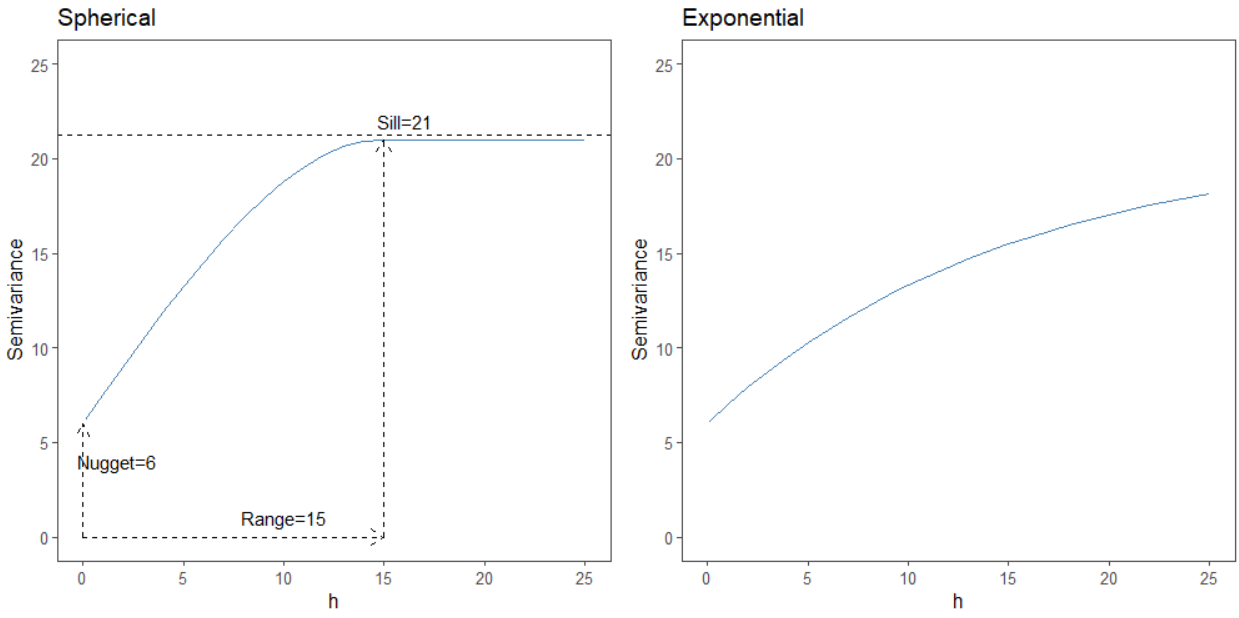


Figure 6: Example semivariograms with the same parameters, but different families

3.3 Covariance Structures

In order to use Kriging interpolation, a covariance structure that describes the spatio-temporal dependencies between any two given points in the domain must be specified. This structure must be positive semi-definite (p.s.d.) in order for the Kriging variances to be valid (i.e. non-negative). A given matrix C is p.s.d. if:

$$x^T C x \geq 0 \quad \forall x \in \mathbb{R}^n$$

The following covariance structures use marginal spatial, temporal, and/or joint spatio-temporal covariance matrices to create valid p.s.d. full covariance structures.

3.3.1 Separable Models

The most simplistic of spatio-temporal covariance structures is the separable covariance model. Its simplicity is in its assumption that the spatial and temporal covariances are separate and independent, making the overall covariance a simple product of the two:

$$C_{Sep}(h, \tau) = C_s(h) * C_t(\tau)$$

The corresponding separable semivariogram is then:

$$\gamma_{Sep}(h, \tau) = sill_{s,t} * (\gamma_s(h) + \gamma_t(\tau) - \gamma_s(h)\gamma_t(\tau))$$

3.3.2 Product-sum Models

The Product-sum model adds on to the separable model by including a sum of the separate covariance structures to the overall covariance, while also adding a weighting parameter k to balance the product and sums. Its covariance and semivariogram formulations are as follows:

$$C_{ProdSum}(h, \tau) = k * C_s(h)C_t(\tau) + C_s(h) + C_t(\tau)$$
$$\gamma_{ProdSum}(h, \tau) = (sill_s * k + 1)\gamma_t(\tau) + (sill_t * k + 1)\gamma_s(h) - k\gamma_s(h)\gamma_t(\tau)$$

3.3.3 Metric Models

Metric models attempt to treat both spatial and temporal lag on the same scale, with both sharing the same covariance structure. This requires an additional term considered the *spatial anisotropy correction* (or *stAni*) to translate the units of the time lag to spatial units. Effectively, this addresses the geometric anisotropy of the time dimension having different scales (and covariances) to the spatial dimension.

$$C_{Met}(h, \tau) = C_{joint}(\sqrt{h^2 + (stAni * \tau)^2})$$

$$\gamma_{Met}(h, \tau) = \gamma_{joint}(\sqrt{h^2 + (stAni * \tau)^2})$$

3.3.4 Sum-metric Models

The Sum-metric model builds on the Metric model, while also including marginal models for both space and time. This adds another layer of complexity over the Metric, and Product-sum structures.

$$C_{SumMet}(h, \tau) = C_{joint}(\sqrt{h^2 + (stAni * \tau)^2}) + C_s(h) + C_t(\tau)$$
$$\gamma_{SumMet}(h, \tau) = \gamma_{joint}(\sqrt{h^2 + (stAni * \tau)^2}) + \gamma_s(h) + \gamma_t(\tau)$$

3.3.5 Simple Sum-metric Models

An extension on the Sum-metric Model is the *Simple Sum Metric* model. This has the same covariance formulation as the Sum-metric structure, but restricts the marginal and joint variograms in the Sum-metric formulation to have no nugget effect. In its place, there is a sole nugget effect on the full variogram.

$$\gamma_{SimSumMet}(h,\tau) = nugget * I[h > 0 \ U \ \tau > 0] + \gamma_{joint}(\sqrt{h^2 + (stAni * \tau)^2}) + \gamma_s(h) + \gamma_t(\tau)$$

3.4 Kriging Types

The main idea of geostatistical interpolation is to optimally weight and combine observed readings to predict a value at some unobserved location on the domain. Kriging interpolation follows Tobler's first law of geography that points nearer to the prediction location should have greater weights.

The Kriging predictor utilizes the observable correlations across spatiotemporal lags to create its weights. Under certain conditions, these weights lead to predictions that are both unbiased and optimal for minimizing the prediction variance, making Kriging the Best Linear Unbiased Predictor (BLUP).

The following predictors are common applications of Kriging, with differences in their mean formulation distinguishing their use cases.

3.4.1 Simple Kriging

In circumstances when the mean value over the domain is known or can be assumed a priori, simple Kriging (SK) can be used for prediction.

This Kriging assumes the following:

- The process Z(s,t) is second-order stationary
- The mean of Z(s,t), μ, must be known

Defining $c_0 = Cov(Z(s_0,t_0), Z(s,t))$ and C_z as the covariance of Z, we can formulate the predictor as:

$$\hat{Z}(s_0,t_0)_{sk} = \mu + \boldsymbol{c_0}^T \boldsymbol{C_z}^{-1}(\boldsymbol{Z} - \mu\boldsymbol{1})$$

Prediction variance of the simple Kriging is then, with the additional term $c_{0,0} = Var(Z(s_0,t_0))$:

$$\sigma^2_{Z,sk}(s_0,t_0) = c_{0,0} - \boldsymbol{c_0}^T \boldsymbol{C_z}^{-1} \boldsymbol{c_0}$$

The process Z follows a multivariate normal likelihood:

$$L(C_z(\theta); Z) = (2\pi)^{-n/2} |\boldsymbol{C}_z(\theta)|^{-1/2} exp(-\frac{1}{2}(\boldsymbol{Z} - \mu\boldsymbol{1})^T (\boldsymbol{C}_z(\theta))^{-1} (\boldsymbol{Z} - \mu\boldsymbol{1}))$$

The maximum likelihood estimate of $\boldsymbol{C}_z$ leads to the optimal simple Kriging predictor.

3.4.2 Ordinary Kriging – Unknown, Constant Mean

In a more common scenario, the mean of a process is assumed to be constant, but is not known at the modelling stage. The appropriate option in this circumstance is ordinary Kriging, which has weaker assumptions:

- The process Z(s,t) is intrinsically stationary
- The mean of Z(s,t), μ is constant across the domain, but not known

The ordinary Kriging predictor is formulated as:

$$\hat{Z}(s_0, t_0)_{ok} = \hat{\mu} + \boldsymbol{c_0}^T \boldsymbol{C}_z^{-1}(\boldsymbol{Z} - \hat{\mu}\boldsymbol{1})$$

and its variance is:

$$\sigma^2_{Z,ok} = (s_0, t_0) = c_{0,0} - \boldsymbol{c_0}^T \boldsymbol{C}_z^{-1}\boldsymbol{c_0} + k$$

with the additional k term representing the uncertainty of estimating the mean:

$$k = (1 - \boldsymbol{1}^T \boldsymbol{C}_z^{-1}\boldsymbol{c}_0)^T (\boldsymbol{1}^T \boldsymbol{C}_z^{-1}\boldsymbol{1})^{-1}(1 - \boldsymbol{1}^T \boldsymbol{C}_z^{-1}\boldsymbol{c}_0)$$

With an unknown mean, the likelihood of the data also relies on an estimated mean μ:

$$L(C_z(\theta), \mu; Z) = (2\pi)^{-n/2}|\boldsymbol{C}_z(\theta)|^{-1/2} exp(-\frac{1}{2}(\boldsymbol{Z} - \mu\boldsymbol{1})^T(\boldsymbol{C}_z(\theta))^{-1}(\boldsymbol{Z} - \mu\boldsymbol{1}))$$

In addition to finding the MLE of $\boldsymbol{C}_z$, the optimal ordinary Kriging predictor requires finding the MLE of μ at the same time. An alternative is to consider the generalized least squares estimate of μ, which limits the likelihood optimization parameter set to just those involved in the covariance structure:

$$\hat{\mu}_{gls} = (\boldsymbol{1}^T \boldsymbol{C}_z(\theta)^{-1}\boldsymbol{1})^{-1}\boldsymbol{1}^T \boldsymbol{C}_z(\theta)^{-1}\boldsymbol{1}\boldsymbol{Z}$$

3.4.3 Kriging with Drift – Unknown, Varying Mean

The assumption of a constant mean across the domain is unrealistic for most spatio-temporal processes. For example, it is natural to propose that the mean particulate matter reading varies based on its spatio-temporal location, if not with the presence of other covariates exogenous to the domain.

Requisite assumptions for Kriging with drift are summarised as follows:

- The process Z(s,t) can be formulated as a deterministic trend function $\mu(s,t)$ and a random residual function $\epsilon(s,t)$

$$Z(s,t) = \mu(s,t) + \epsilon(s,t)$$

- Mean function $\mu(s,t)$ is assumed to be a linear combination of predictors that are known at all locations (s, t) on the domain

$$\mu(s,t) = \sum_i \beta_i x_i(s,t)$$

  with vector of coefficients $\boldsymbol{\beta}$ being unknown

- Residual function $\epsilon(s,t)$ is assumed to be intrinsically stationary with a mean of zero

When the deterministic mean is solely a function of the location on the domain, the interpolation is known as *Universal Kriging* (UK). If instead, the deterministic mean is a function of other feature-space predictors, this is considered to be *Kriging with External Drift* (KED).

Regardless of which set of covariates are used, the Kriging with drift predictor is formulated as:

$$\hat{Z}(s_0,t_0)_{ked} = \boldsymbol{x}(s_0,t_0)^T\hat{\boldsymbol{\beta}} + \boldsymbol{c_0}^T\boldsymbol{C}_z^{-1}(\boldsymbol{Z} - \boldsymbol{X}^T\hat{\boldsymbol{\beta}})$$

and its variance is:

$$\sigma^2_{Z,ked} = (s_0,t_0) = c_{0,0} - \boldsymbol{c_0}^T\boldsymbol{C}_z^{-1}\boldsymbol{c_0} + k$$

with the k term being:

$$k = (\boldsymbol{x}(s_0,t_0) - \boldsymbol{X}^T\boldsymbol{C}_z^{-1}\boldsymbol{c}_0)^T(\boldsymbol{X}^T\boldsymbol{C}_z^{-1}\boldsymbol{X})^{-1}(\boldsymbol{x}(s_0,t_0) - \boldsymbol{X}^T\boldsymbol{C}_z^{-1}\boldsymbol{c}_0)$$

Maximum likelihood estimates for the covariance matrix $\boldsymbol{C}_z$ and mean function parameters $\boldsymbol{\beta}$ are found simultaneously for the optimal prediction, similarly to Ordinary Kriging. Again, the generalized least squares estimator of $\boldsymbol{\beta}$ can be used instead for profile likelihood maximization which limits the number of parameters to be optimized.

$$\hat{\boldsymbol{\beta}}_{gls} = (\boldsymbol{X}^T\boldsymbol{C}_z(\theta)^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{C}_z(\theta)^{-1}\boldsymbol{Z}$$

3.4.4 Residual Kriging – Separate Drift Estimation and Residual Interpolation

An alternative to Kriging with External Drift is *Regression* or *Residual Kriging* (RK). Instead of estimating the trend in the Kriging process, Regression Kriging first de-trends the data using an estimated mean function and then applies simple (with assumed mean of 0) or ordinary Kriging to interpolate the residuals.

Kriged residuals and the separately predicted mean are then combined to form the predictor for RK:

$$\hat{Z}(s_0,t_0)_{rk} = \hat{\mu}(s_0,t_0) + \hat{\epsilon}_{ok}(s_0,t_0)$$

The variance term is then a sum of mean prediction's variance, the variance of the simple or ordinary Kriging prediction used to interpolate the residuals, and the covariance between the two.

$$\sigma^2_{rk}(s_0, t_0) = Var(\hat{\mu}_{rk}(s_0, t_0)) + Var(\hat{\epsilon}_{ok}(s_0, t_0)) + 2Cov(\hat{\mu}_{rk}(s_0, t_0), \hat{\epsilon}_{ok}(s_0, t_0))$$

When the mean function is estimated using OLS, the covariance between the predicted mean and the residuals is assumed to be zero, due to the orthogonality between the two. This eliminates the covariance term, making the variance formulation much simpler. The covariance term is less easily ignored for other forms of mean estimation.

Performing regression Kriging requires an additional step to the other Kriging methods featuring drifts, but it comes at the benefit of allowing nonlinear trend models and neighborhood-level interpolation to be considered.

3.5 Kriging with Large Datasets

Difficulties arise for spatio-temporal model fitting and prediction as the size of the dataset increases. Matrix decomposition for the covariance inversion necessary in both maximum likelihood estimation and in creation of Kriging weights comes at a computational cost of $O(n^3)$ [].

Due to the scale of our data, with roughly 366 daily observations of 100 training sites, a different approach for model fitting and prediction is required.

3.5.1 Model Fitting Through the Semivariogram

A common light-weight approach to modeling the spatio-temporal dependency of a random field in geostatistics is to consider the fitted covariance matrix through the fitted semivariogram.

Assuming isotropy and intrinsic stationarity, we can consider the semivariogram to be only a function of spatial and temporal distance h &τ. Then we can compute the *empirical semivariogram* using Matheron's method-of-moments estimator:

$$\gamma_{emp}(h, \tau) = \frac{1}{2|N(h, \tau)|} \sum_{i}^{N(h,\tau)} (Z(s_i, t_i) - Z(s_i + h, t_i + \tau))^2$$

Here $|N(h, \tau)|$ is the number of sample pairs within a spatio-temporal lag of h and h. These lags are generally binned to reduce the noise of the semivariogram's structure.

For simple and ordinary Kriging, the empirical semivariogram is fitted on the observed data Z, whereas for Kriging with external drift, universal Kriging, or regression Kriging, the semivariogram is fitted on the residuals of the observed data after de-trending.

Proposed theoretical semivariogram models are evaluated against the empirical semivariogram, with the best fitting theoretical model determined as the one that minimizes the weighted least squares difference in their binned semivariance values:

$$Error = \sum_{j}^{n} w_j(\gamma_{emp}(h_j, \tau_j) - \hat{\gamma}(h_j, \tau_j))^2$$

Various weighting schemes are used, including equal weights for all bins. The weighting scheme we use incorporates the empirical γ value, as well as the number of observation pairs (N_j) that fall into bin j. This weighting scheme prioritizes lower errors on semivariogram bins that many observation pairs fall into, as well as prioritizing bins with smaller lags, which should have a greater impact on predictions given their proximity.

$$w_j = \frac{N_j}{\gamma_{emp}(h_j, \tau_j)^2}$$

Assuming second-order stationarity, the estimated theoretical semivariogram can be used to retrieve the covariance. This in part can calculate the generalized least-squares estimates of the mean μ or vector of trend parameters $\boldsymbol{\beta}$ for ordinary and Kriging with drift respectively. The latter terms can then be used to produce a set of genearlized least-square residuals of the data, and a new semivariogram can be fitted for Kriging with drift. In turn, this can re-generate the $\boldsymbol{\beta}$ parameters, which can then create a new residual semivariogram. This process can repeat indefinitely, although simulations have shown this to be near-optimal after 1 iteration. Full optimality requires the covariance to be fitted directly, rather than the semivariogram.

3.5.2 Local Neighborhood Kriging

Local Kriging limits the field of observed data used in interpolation to a set number k. An appropriate k limits the samples used to just those that would fall within the range of the modeled semivariogram. This would imply that the values used in interpolation are the ones that have meaningful spatio-temporal correlation with the prediction site.

One benefit of local Kriging is the decrease in computational complexity, as the covariance matrix used in interpolation is limited to only the k selected observations. The second benefit comes in the relaxation of global mean stationarity, as the mean would only need to be constant within each neighborhood.

The formulas for local Kriging are the same as for global Kriging, only we use a reduced set of the sample data, $\tilde{\boldsymbol{Z}}$, which we define as the k-nearest samples to the prediction location (s_0, t_0).

$$\hat{Z}(s_0, t_0)_{local\ ok} = \hat{\mu} + \boldsymbol{c_0}^T \boldsymbol{C}_{\tilde{z}}^{-1}(\tilde{\boldsymbol{Z}} - \hat{\mu}\boldsymbol{1})$$

This reduced set disrupts the cycle of Kriging with drift, unless localized semivariograms are also fit to the neighborhood samples. As noted in the previous section, the global estimated drift informs the empirical semivariogram, which in turn informs the estimated drift parameters used in the prediction. If we localize the drift parameter estimation, we will only use a subset of the covariate values that were used to estimate the initial global trend and semivariogram. This could possibly lead to decent predictions, but the semivariogram-based Kriging system will no longer be near-optimal.

3.6 Model Evaluation

To evaluate the predictive capability of our models, we look at the following metrics:

- Root Mean Squared Prediction Error (RMSPE) serves as a measure of general prediction accuracy

$$RMSPE = \sqrt{\frac{\sum_i^n ((\hat{Z}(s_i, t_i) - Z(s_i, t_i))^2}{n}}$$

- Mean Prediction Error (MPE) observes the average bias of the predictions:

$$MPE = \frac{\sum_i^n \hat{Z}(s_i, t_i) - Z(s_i, t_i)}{n}$$

- Mean Prediction Standard Error (MPSE) gauges the average uncertainty in the model's predictions:

$$MPSE = \frac{\sum_i^n \hat{\sigma}^2_{Kriging}(s_i, t_i)}{n}$$

While the modeling is performed on log $PM_{2.5}$, it is more important to predict actual values of particulate matter. Thus, while we train our models on the transformed response, our final evaluation on model predictive capability will be done on back-transformed predictions. Simple transformation to the original scale introduces bias, generally under-predicting the value ([2]). Therefore, we utilize corrected back-transformations.

For Kriging, the corrected, back-transformed prediction ([5]) is formulated as:

$$\hat{Z}(s_0, t_0)_{bt} = e^{(Z(s_0,t_0)+\sigma(s_0,t_0)^2/2)}$$

The Kriging variance follows as:

$$\sigma(s_0, t_0)^2_{bt} = (e^{(\sigma(s_0,t_0)^2)} - 1) * e^{(2\hat{Z}(s_0,t_0)+\sigma(s_0,t_0)^2)}$$

4 Modeling Process

For our analysis, we assume that the our log-transformed response, $PM_{2.5}$, is spatially isotropic and second-order stationary (with a locally stationary, unknown mean) *after having been de-trended by an deterministic linear trend model.* The coefficients for our trend function will be determined by ordinary least squares:

$$\hat{\beta}_{ols} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Z}$$

In addition, we assume that the predicted trend and the kriged residuals of the response are orthogonal. These are the base assumptions for our Residual Kriging model.

4.1 Holdout Set Creation

To validate our modeling process, modeling assumptions, and to compare final Kriging predictions, we evaluate our predictors against a test set. Twenty-six sensor locations are omitted from the full set to demonstrate the accuracy of the models on "unobserved" locations that had no influence on the fitting of the spatio-temporal models. This leaves an even one hundred locations to use as a training set, for a roughly eighty-twenty training-testing split.

4.2 Land-use Regression Structure

4.2.1 Hierarchical Lasso

A capable model for de-trending the data requires some complexity, as the generation of air pollution comes from a multitude of sources. In the interest of adding substantial predictive capability to our estimated trend function, we consider all main effects and all pairwise interactions of our 25 possible covariates. As this is a substantial set of possible covariates, we use a hierarchical group lasso regression to determine a suitable subset of predictors. This regularization technique is performed through the R package *glinternet*, and differs from traditional lasso as it requires the two involved main effects be added to the model whenever the lasso deems a pairwise interaction to be non-zero [4].

To ensure proposed predictor subsets are predictive, we perform the lasso regularization in a 10-fold cross-validation procedure. The steps are as follows:

1. The 100 training locations – and their respective daily observations – are randomly split into 10 groups of 10.

2. Nine groups are fit with the hierarchical group lasso regression repeatedly for 50 possible penalty terms ranging from high ($\lambda = 0.001$) to low regularization ($\lambda = 0.00001$).

3. Each of the fifty fitted lasso regressions is used to predict the held-out tenth group, and prediction accuracy on the log scale is recorded.

4. Steps 2 and 3 are repeated with each of the ten groups being held out as the validation group once.

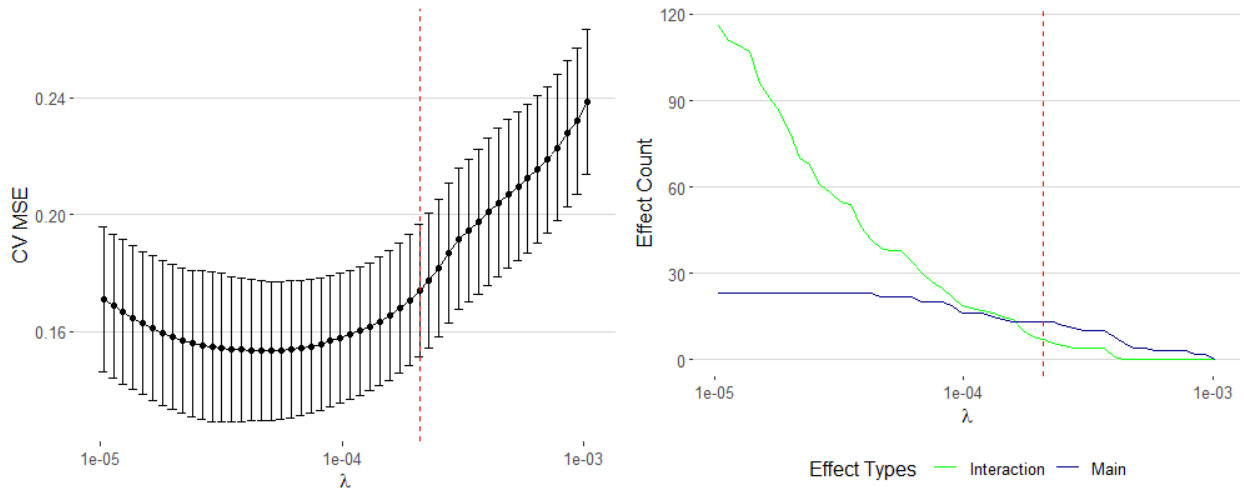


Figure 7: 10-fold cross-validation results for covariate subset selection

Results of the cross-validation in Figure 7 show higher prediction accuracy for values of λ below 0.0001. This, however, corresponds to a glut of interaction terms. For a more parsimonious model, we instead opt for a λ of approximately 0.0002 – this is a cutoff that includes 20 total effects (7 being interactions). With far fewer parameters, this still manages to be within one cross-validation prediction error standard error from the interaction heavy model that minimizes the prediction error.

4.2.2 Fitted Model Summary

	Log(PM2.5)		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	-10.066	-11.827 – -8.304	**<0.001**
elevation	-0.047	-0.053 – -0.042	**<0.001**
ndvi	0.612	0.531 – 0.692	**<0.001**
temp	0.980	0.832 – 1.129	**<0.001**
pressure	0.013	0.011 – 0.014	**<0.001**
wind	-0.153	-0.162 – -0.143	**<0.001**
aod47 imp	1.151	1.101 – 1.200	**<0.001**
precipitation	-0.034	-0.038 – -0.031	**<0.001**
crops prop	-0.272	-0.312 – -0.232	**<0.001**
built prop	0.011	-0.035 – 0.056	0.646
pop density	0.000	0.000 – 0.000	**<0.001**
trees prop	0.739	0.616 – 0.862	**<0.001**
humidity	0.002	0.001 – 0.004	**<0.001**
month [2]	-0.217	-0.247 – -0.188	**<0.001**
month [3]	-0.193	-0.225 – -0.162	**<0.001**
month [4]	-0.235	-0.271 – -0.200	**<0.001**
month [5]	-0.603	-0.639 – -0.568	**<0.001**
month [6]	-0.820	-0.857 – -0.782	**<0.001**
month [7]	-0.812	-0.852 – -0.773	**<0.001**
month [8]	-0.882	-0.922 – -0.841	**<0.001**
month [9]	-0.655	-0.693 – -0.618	**<0.001**
month [10]	-0.551	-0.584 – -0.519	**<0.001**
month [11]	0.023	-0.007 – 0.052	0.135
month [12]	-0.159	-0.192 – -0.126	**<0.001**
ndvi * trees prop	-2.236	-2.445 – -2.026	**<0.001**
temp * precipitation	0.002	0.002 – 0.002	**<0.001**
temp * pressure	-0.001	-0.001 – -0.001	**<0.001**
temp * wind	0.002	0.002 – 0.003	**<0.001**
temp * humidity	-0.001	-0.001 – -0.001	**<0.001**
Observations	35373		
R^2 / R^2 adjusted	0.376 / 0.376		

Figure 8: Estimated linear trend model summary

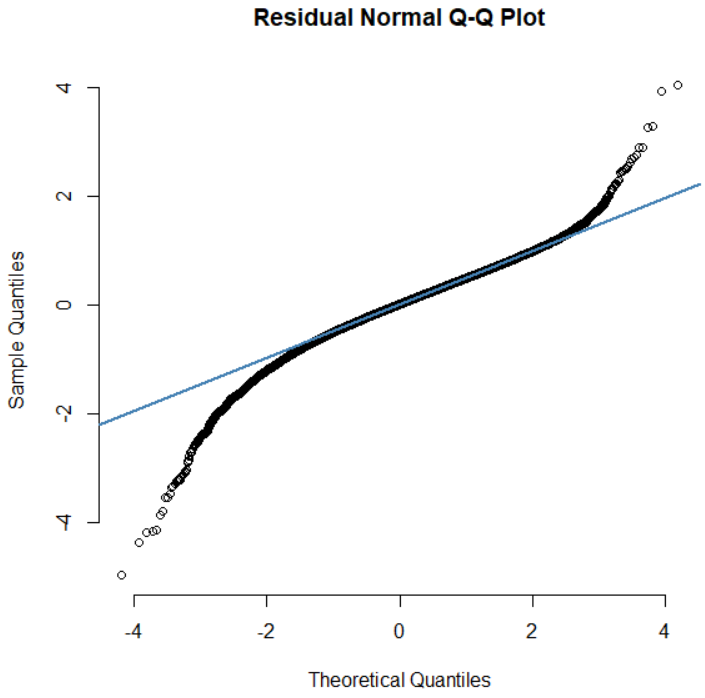
The fitted estimates of our land-use regression model are largely intuitive. In relation to our reference month of January 2020, most other months are associated with lesser pollution readings. Elevation, wind, and precipitation are all sensibly fitted with negative effects. Other environmental factors, like temperature, tree proportion, and NDVI all have counter-intuitive main effects. However, temperature's association is muddled by four interactions with other meteorological factors, and the NDVI*Tree interaction indicates that forested areas are associated negatively with pollution when the NDVI value is high (ie the trees are alive). The variance explained by this model is not remarkable, however, the goal is to reduce the variance that cannot be explained by the semivariogram in the Kriging process as much as possible. Going forward, we use this model to de-trend the training and testing set in the regression Kriging setting.

4.2.3 De-trended Residuals

The bulk of the residuals follow the normal distribution, with exceptions coming in the large tails at either end. A number of large, positive residuals come from the more anomalous pollutant events that are triggered by phenomena not covered in the model. While the model is able to create general explanations for gradual increases in pollution, it has no useful covariates that capture why air pollution spikes.

Higher negative residuals are largely a result of the log-transformation of the response. While this brings the overall distribution to a more normal setting, it does expand the variation in low levels of air pollution. Readings of $PM_{2.5}$ between 0 and 3 are near-negligible differences in air quality. However, once log-transformed, these low volumes cover nearly half the range of the distribution. Thus, the model, which is unable to explain these somewhat manufactured variances between minimal levels of air pollution, results in a good deal of over-estimation.

The more important residuals to consider are the under-estimations, as they will be amplified in the back-transformation process. We move forward with the Kriging prediction, with the concept that interpolation of the sample residuals will lead to better predictions.



(a) Normal Q-Q Plot of the de-trended data residuals

4.3 Spatio-Temporal Dependence

4.3.1 Checking for Anisotropy

In our process, we assume isotropy to simplify the form of our semivariogram and spatio-temporal dependence modeling. We can make a rough assessment of the validity of this assumption by averaging the daily spatial directional variograms across the full year, weighting points by the number of sample pairs at each distance, each day.

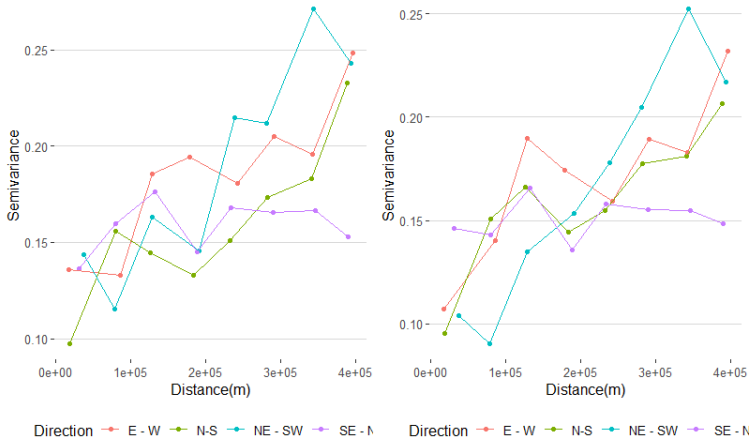


Figure 9: Averaged daily spatial directional semivariograms of data (left) and residuals(right)

Variance in the intermediary steps aside, all directions but the SE-NW direction have a similar semivariance path in the original data, with the N-S direction having a lower nugget. The de-trending corrects this, aligning the nuggets of those three semivariogram lines. However, the issue persists between both estimates of the variogram line for the SE-NW direction being flat. Assuming isotropy would lead to incorrect weights for sample points to the south east or north west of our prediction point. This is just one form of isotropy (spatial), with spatio-temporal anisotropy being possible in all three-dimensional directions of the domain. We will continue to assume isotropy, and determine if any violations are acceptable based on the success of our predictions.

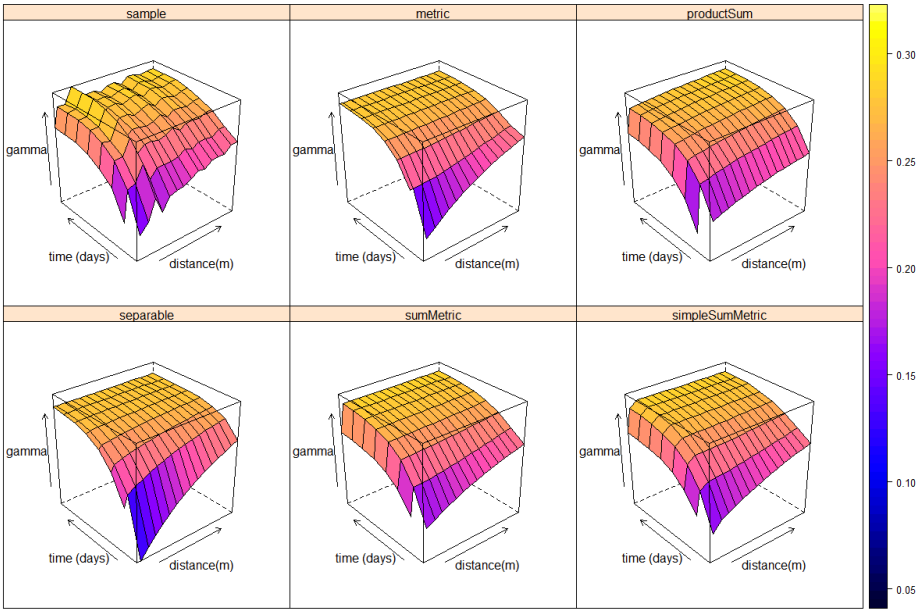
4.3.2 Semivariogram Modeling

When modeling the spatio-temporal dependence, we consider all semi-variogram structures outlined in 3.3. For each, we further consider all combinations of the exponential and spherical models for the marginal spatial, marginal temporal, and joint semi-variograms. These theoretical models are evaluated against the empirical semivariogram using weighted least squares error as described in Section 3.5.1. Weighted mean squared errors for the fits are shown in Table 2, with the best fitting combination for each spatio-temporal semivariogram in bold.

Table 2: Theoretical semi-variogram weighted mean squared error fits (Marginal structure choices displayed as Spatial+Temporal)

Model	Joint	Exp+Exp	Exp+Sph	Sph+Exp	Sph+Sph
Metric	-	**229.89**	-	-	375.83
Product-Sum	-	**189.08**	811.15	207.42	817.52
Separable	-	**634.78**	1292.07	953.74	933.96
Sum-Metric	Exp	101.33	**100.67**	101.35	101.25
	Sph	107.60	104.02	109.03	105.51
Simple	Exp	186.82	185.74	189.89	**100.08**
Sum-Metric	Sph	191.40	186.21	194.48	190.22

A 3-dimensional visualization of the empirical and theoretical semivariograms is shown in Figure 14. Similar to the results found by weighted mean squared error, the best visual fit of the empirical semivariogram belongs to the sum-metric and the simple sum-metric models.



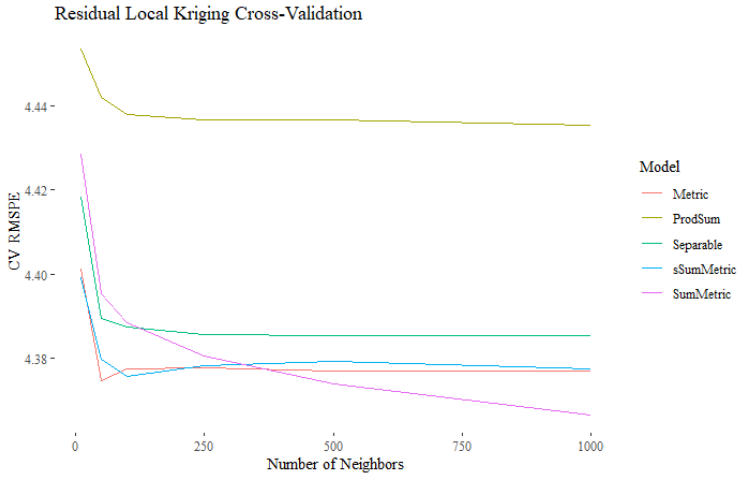
(a) Residual semivariograms – empirical (top-left) and theoretical

Figure 10

4.4 Model Selection

We choose between the best-fitting parameterizations of each of the five semivariogram model types, as well as the optimal number of samples to include in local Kriging, by revisiting the cross-validation folds used in section 4.2. For each step of the process, the nine training folds

are used as samples to predict the held-out fold via residual Kriging with the best-fitting semivariograms for each semivariogram family used in the weighting formulation. As we now have Kriging variances, we aim for minimization of back-transformed prediction accuracy, rather than accuracy on the log scale, as was optimized for in section 4.2.



Results in figure 4.4 show marginal differences in neighborhood choices in all models outside of a clear increase in accuracy when moving from ten to fifty neighbors. The optimal spatio-temporal model for our residual Kriging is the sum-metric model, which matches both the visual test, and the WLS scores noted in the previous section. This model is composed of a joint exponential semivariogram, a marginal spatial exponential semivariogram, and a marginal temporal spherical semivariogram. Additionally, the optimal neighborhood size for prediction is the one thousand nearest neighbors.

5 Results

For our final predictions, we use our fitted sum-metric semivariogram from Section 4.3 to perform residual Kriging for all daily values of the 26 locations in our testing set. The Kriging interpolates using the de-trended observations of the nearest 1000 training samples to each prediction location, per our cross-validation findings. Results from this Kriging are then added to the predictions made for the test set by our linear trend from Section 4.2 to get our final log $PM_{2.5}$ prediction set.

As a baseline comparison, we also make predictions using just the estimated linear trend from Section 4.2, as well as predictions using an ordinary Kriging model optimized in Section A.1.

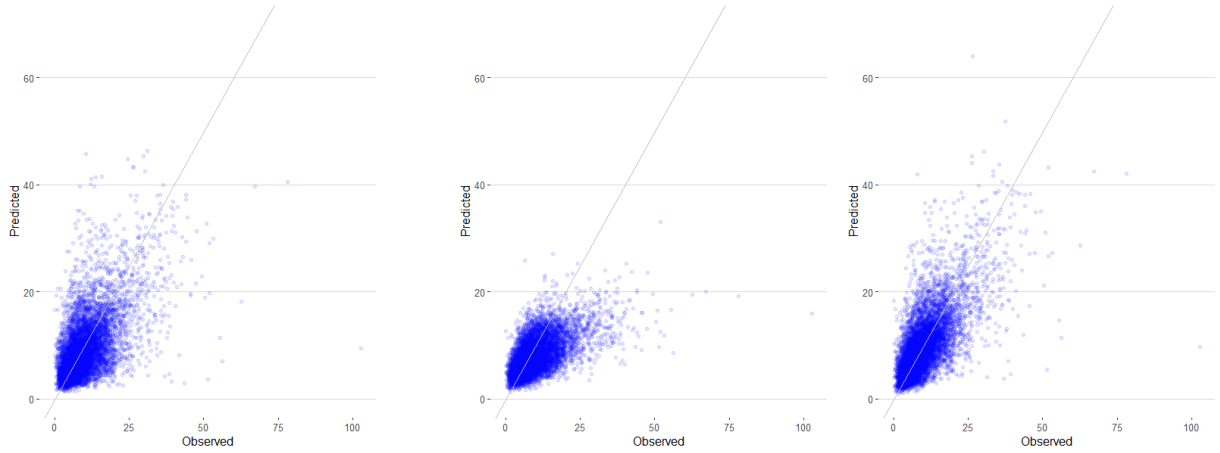
5.1 Hold-out Set Predictions

After predicting the log $PM_{2.5}$ at each location, point estimates and prediction variances are back-transformed to standard scale using the equations in section 3.6.

Table 3: Hold-out set prediction metrics comparison

Predictor	RMSPE	MPE	MPSE	COR
Mean Trend Only	5.597	-1.833	3.229	0.605
Ordinary Kriging	5.591	-0.623	3.785	0.590
Residual Kriging	4.956	-0.208	3.734	0.686

All back-transformed prediction metrics prefer the residual Kriging over ordinary Kriging, or the simple land-use regression predictions. The combined method of de-treneding and residual interpolation leads to lower RMSPE, paired with a marginally lower average prediction standard error (the LUR prediction standard error was omitted as its assumption of independent residuals is clearly violated by the spatio-temporal dependencies here).



(a) Ordinary Kriging (b) OLS - No Kriging (c) Residual Kriging

Figure 11: Back-transformed model predictions

The benefit of de-trending the data before Kriging can be observed by comparing the plots of predicted vs observed values in Figure 11. Removing some of the variance in the dataset using our covariate set leads to a tighter cone of predictions. Predictions from the land-use regression without Kriging have a smaller spread, but a much worse bias. Interpolating the residuals of these predictions goes some way in correcting this, as is evident in the plots.

5.2 Full Grid Predictions

5.2.1 Ordinary and Residual Kriging Comparison

Interpolations on the full study area illustrate the greater granularity that is achieved in the residual Kriging predictions. Even with the mean free to change within each 1000-sized neighborhood, the ordinary Kriging is not as capable at estimating high-resolution trends.

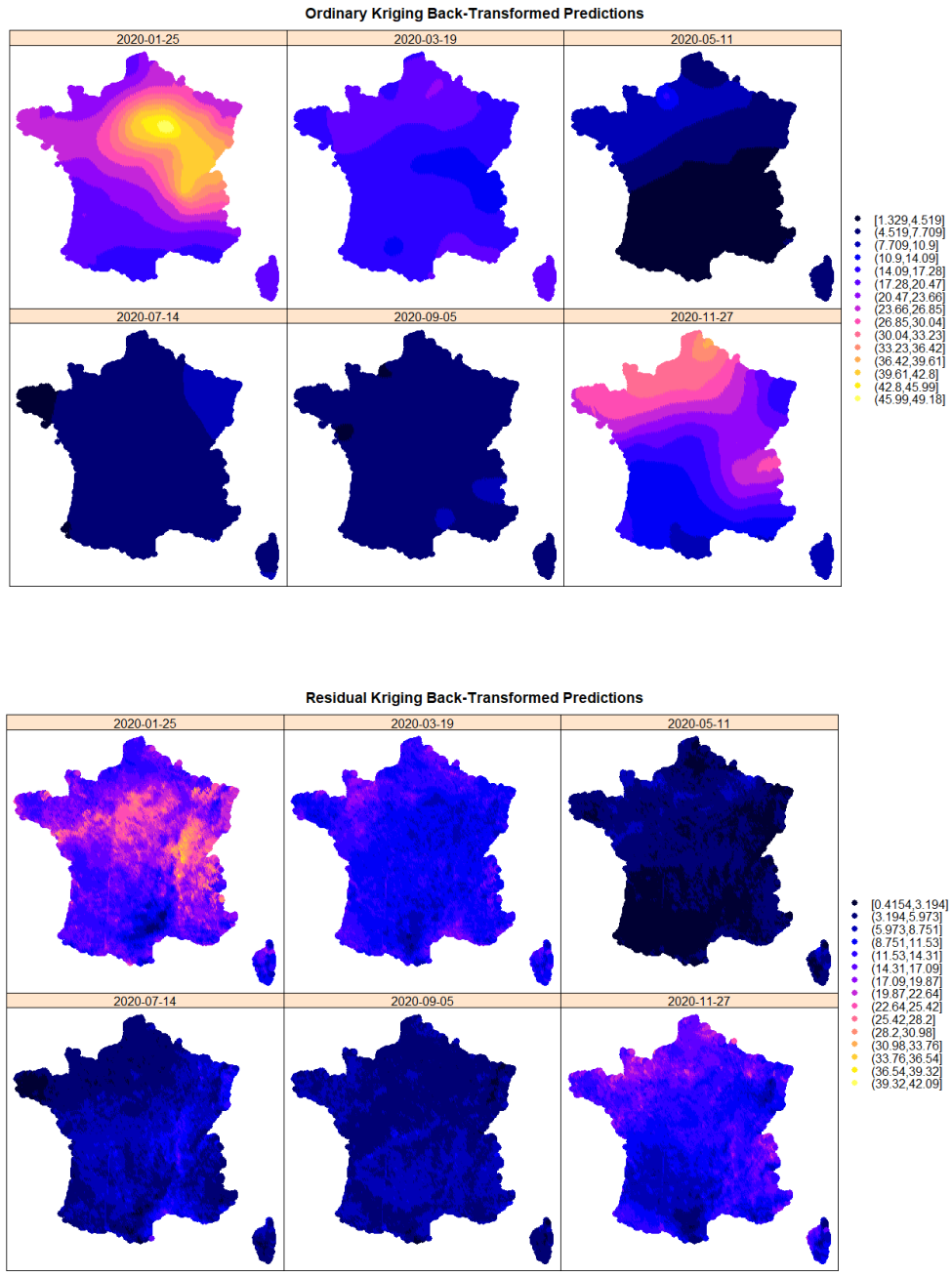


Figure 12: Study Area predictions for 6 days in 2020

5.2.2 Kriging Standard Error Back-Transformation Changes

On the log scale, the estimate Kriging variance and standard deviation are sensibly tied to the location of observed samples, as we are more sure of predictions closer to where we have observations. From day to day, these measures hardly differ. However, once we back-transform, the prediction variance becomes a function of the predicted mean. We no longer see the pattern of observed locations, and instead we see high standard error estimates wherever we see high predictions of pollution.

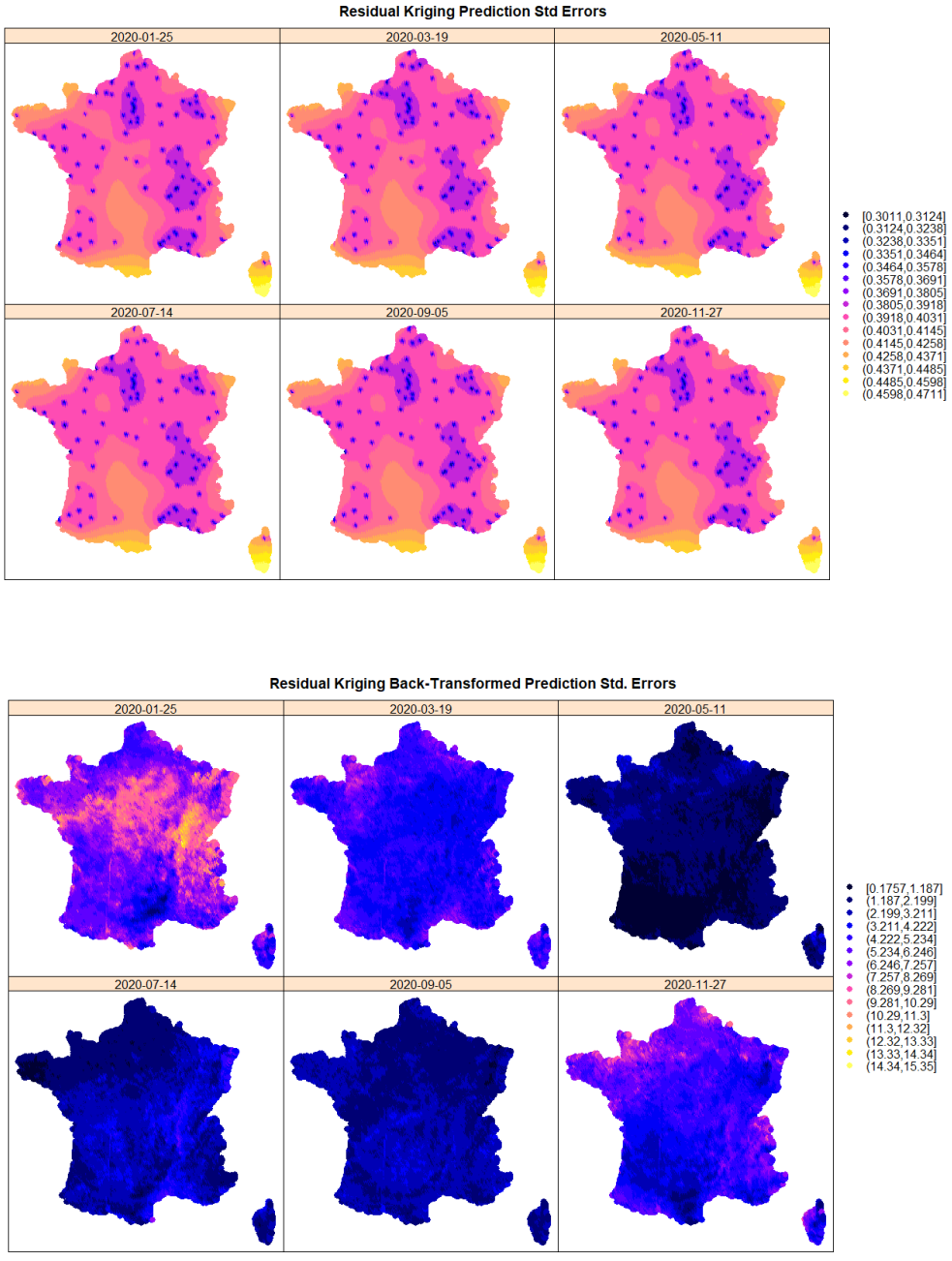


Figure 13: Study Area prediction standard errors

6 Conclusions

Based on the daily average concentration of $PM_{2.5}$ in 2020 in Metropolitan France, we consider the performance of Kriging interpolation from six different spatiotemporal models both with and without mean de-trending. Predictive metrics on both the log scale and the back-transformed scale show that the individual components of land-use regression and ordinary Kriging interpolation can both be improved on by combining the two in residual Kriging. In the de-trending process, meteorological variables, satellite readings, and static variables such as land use and population density are useful in explaining some of the variation in $PM_{2.5}$. Notably, of the included predictor categories, no road covariates are included in the de-trending, either due to redundancy or a lack of direct relation with the response.

Sum-metric structures prove to have the best approximation of the spatio-temporal dependencies of $PM_{2.5}$ based both on fit metrics, and prediction both in and out of sample. Further improvements on the dependency structure, and thus on the predictions, could be had by addressing possible anisotropy in the spatial, temporal, or spatio-temporal dimension. Other improvement in predictive ability could come from more granular predictive variables, or the inclusion of other auxiliary features such as agricultural, industrial, or traffic emissions.

A Appendix

A.1 Ordinary Kriging Model

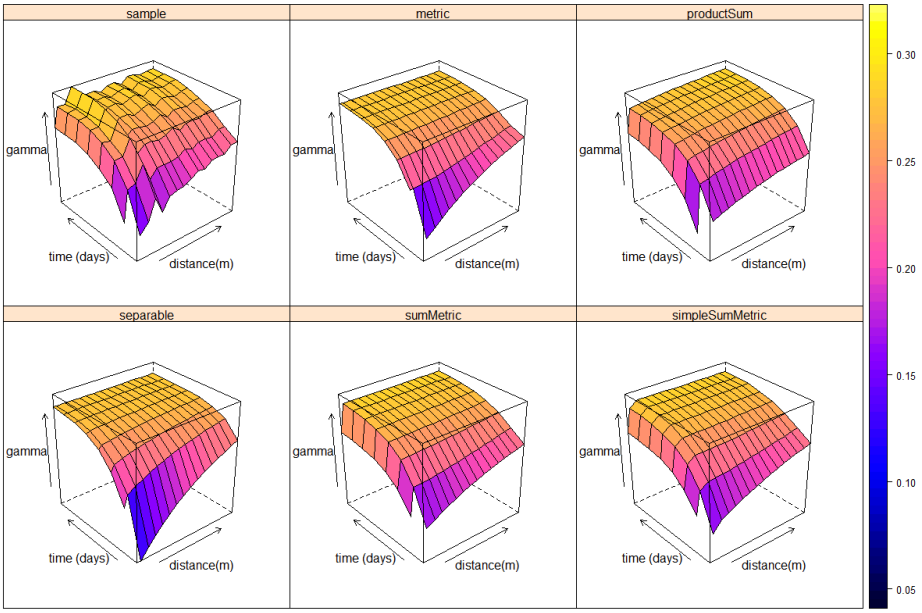
Similar to our process for fitting a spatio-temporal structure for residual Kriging, we also evaluated several models for an ordinary Kriging predictor.

The weighted mean square error fits to the empirical semivariogram are noted in Table 4.

Table 4: Theoretical Variogram fits for Ordinary Kriging (Marginals displayed as Spatial+Temporal)

Model	Joint	Exp+Exp	Exp+Sph	Sph+Exp	Sph+Sph
Metric	-	**210.47**	-	-	394.78
Product-Sum	-	**160.16**	806.33	207.22	765.23
Separable	-	764.81	**541.07**	1312.22	1180.91
Sum-Metric	Exp	196.50	67.66	65.23	**63.62**
	Sph	72.84	66.98	71.80	64.88
Simple	Exp	252.70	148.19	153.10	152.55
Sum-Metric	Sph	151.95	**147.11**	155.78	150.52

Similarly, we visualize the empirical and the best fitting semivariograms.



(a) Residual semivariograms – empirical (top-left) and theoretical

Figure 14

By weighted least squares fit and by visual inspection, the sum-metric model is chosen for the ordinary Kriging model.

A.2 Formulas and Derivations

A.2.1 LCC Coordinate Formulas

Given the sample longitude λ, LCC reference longitude λ_0, sample latitude δ, reference latitude δ_0, radius of the Earth R, and standard parallel latitudes δ_1 and δ_2, the projected coordinates can be described as:

$$\lambda_{lcc} = \rho sin(n(\lambda - \lambda_0))$$
$$\delta_{lcc} = \rho_0 - \rho cos(n(\lambda - \lambda_0))$$

with supporting functions:

$$n = \frac{ln(cos(\delta_1)sec(\delta_2))}{ln(tan(1/4\pi + 1/2\delta_2)cot(1/4\pi + 1/2\delta_1}$$

$$\rho = RFcot^n(1/4\pi + 1/2\delta)$$
$$\rho_0 = RFcot^n(1/4\pi + 1/2\delta_0)$$
$$F = \frac{cos(\delta_1)tan^n(1/4\pi + 1/2\delta_1)}{n}$$

A.2.2 Semi-variogram Relation to Covariance

Under second-order stationarity, the semivariogram can be written as a function of spatiotemporal lags:

$$2\gamma(h,\tau) = Var(Z(s,t) - Z(s+h,t+\tau))$$

Re-writing the variance term as expectations, we have:

$$2\gamma(h,\tau) = E([Z(s,t) - Z(s+h,t+\tau) - E(Z(s,t) - Z(s+h,t+\tau)]^2)$$

Expanding the nested expectation term $E([Z(s,t) - Z(s+h,t+\tau)$ via the linearity of expectation leads to a 0 term, as the second-order stationarity assumption requires the mean to be invariant to translation. Removing this 0 term, we have:

$$2\gamma(h,\tau) = E([Z(s,t) - Z(s+h,t+\tau)]^2)$$

We next add and subtract the mean, μ, and expand the square:

$$2\gamma(h,\tau) = E[(Z(s,t) - \mu)^2 + (Z(s+h,t+\tau) - \mu)^2 - 2(Z(s,t) - \mu)(Z(s+h,t+\tau) - \mu)]$$

Again using linearity of expectation, we re-write the last term as the covariance. Under second-order stationarity, this only depends on the spatio-temporal lags. Further, the first two terms can be re-written as variances:

$$2\gamma(h,\tau) = Var(Z(s,t)) + Var(Z(s+h,t+\tau)) - 2C_z(h,\tau)$$

The variance is invariant to translation under second-order stationarity, allowing the semivariogram to reduce to the following:

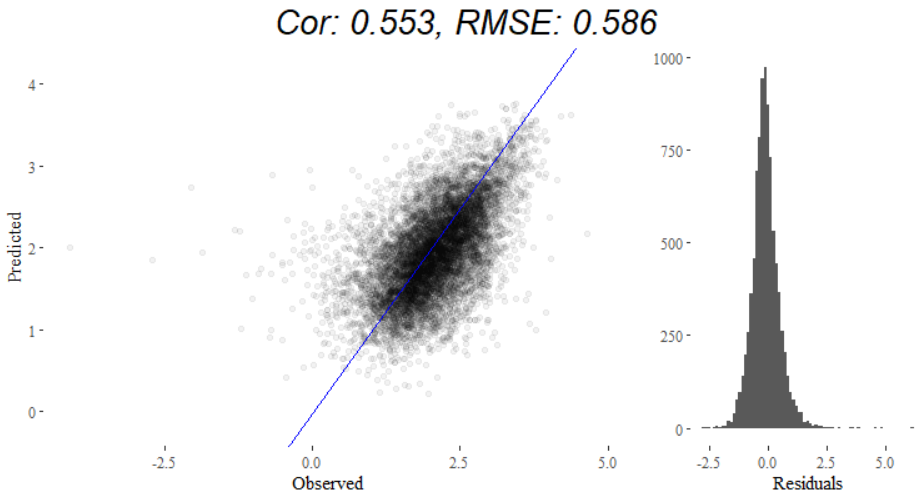
$$\gamma(h,\tau) = \sigma^2 - C(h,\tau)$$

Therefore, we can directly relate the semivariogram to the covariance.

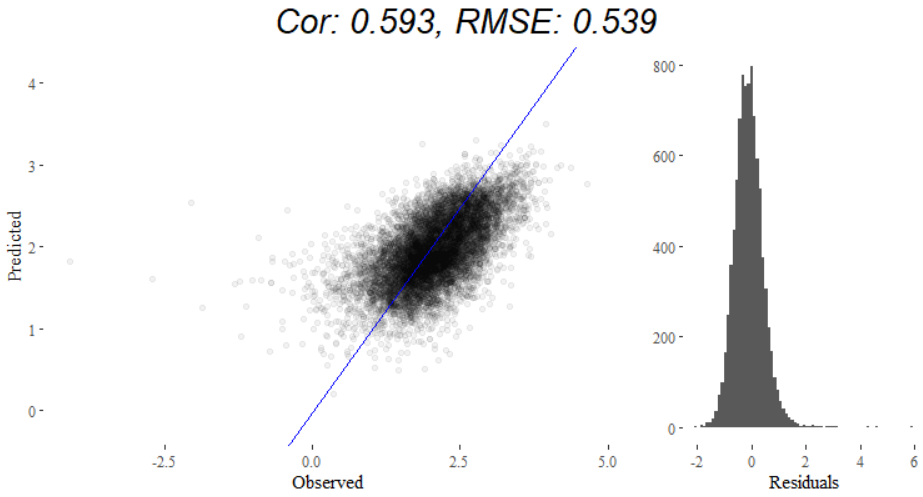
B Additional Visualizations

B.1 Log $PM_{2.5}$ Holdout Error Plots

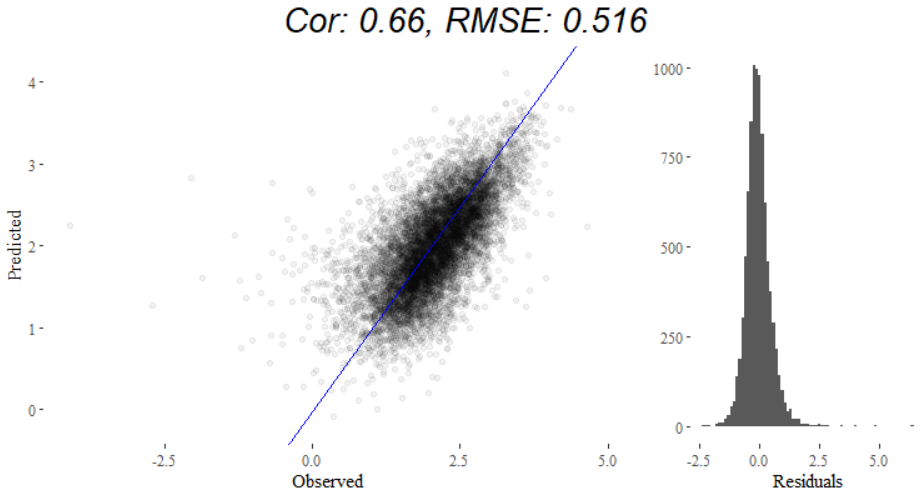
The log transformation, while capable in making the response more normal, also punishes the models for not being able to adequately differentiate exp(-2) from exp(1). This is a negligible difference in actual particulate matter, but a meaningful error in the metrics. All the models suffer relatively equally, as viewable in these plots.



(a) Ordinary Kriging



(b) Linear Trend (No Kriging)



(c) Residual Kriging

Figure 15: Test set predictions

B.2 Study Area Log $PM_{2.5}$ Predictions

Similar to what was mentioned in the results section, the interpolation from ordinary Kriging is more general due to the lack of trend estimation. The results hold on the original log $PM_{2.5}$ predictions.

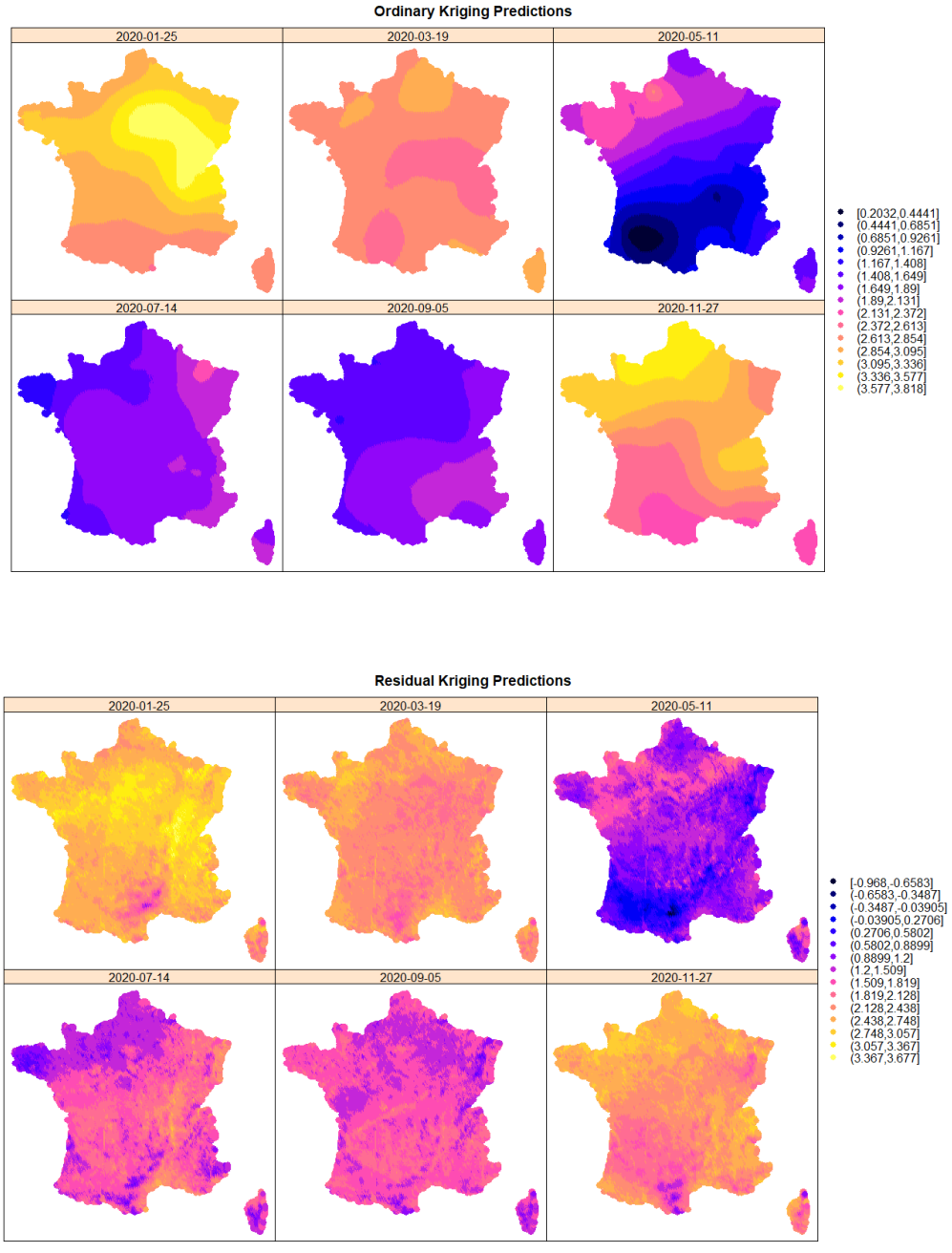


Figure 16

References

[1] Didan, K. (2015). *MYD13A2 MODIS/Aqua Vegetation Indices 16-Day L3 Global 1km SIN Grid V006 [Data set]* . NASA EOSDIS Land Processes DAAC. Accessed 2021-11-23 from https://doi.org/10.5067/MODIS/MYD13A2.006

[2] Cressie N (1993) Statistics for spatial data. 135-138

[3] Wikle, C. K., Zammith-Mangion, A., amp; Cressie, N. (2019). Descriptive Spatio-Temporal Statistical Models. In Spatio-temporal statistics with R. essay, CRC Press.

[4] Lim, M and Hastie, T (2013). *Learning interactions through hierarchical group-lasso regularization*

[5] Denby, Bruce and Schaap, Martijn and Segers, Arjo and Builtjes, Peter and Horálek, Jan (2008). Comparison of two data assimilation methods for assessing PM10 exceedances on the European scale *Atmospheric Environment* vol. 42, 7124–7126

[6] Anenberg, S., Miller, J., Minjares, R. et al. Impacts and mitigation of excess diesel-related NOx emissions in 11 major vehicle markets. *Nature 545*, 467–471 (2017). https://doi.org/10.1038/nature22086

[7] R-PUR. France emits 164,000 tons of $PM_{2.5}$ – Source: European Environment Agency. https://www.r-pur.com/en/blogs/air/france-particules-fines-pm25-europe-pollution-air

[8] Reuters (2020). EU takes France to court for second time over air pollution. https://www.reuters.com/article/us-eu-environment-france-idUKKBN27F1ZY

[9] Hoek, G and Beelen R, and de Hoogh, K and Vienneau, D, and Gulliver, J, and Fischer, P, and Briggs, D (2008). A review of land-use regression models to assess spatial variation of outdoor air pollution. *Atmospheric Environment* 42, 7561-7578

[10] Lei Huang, Can Zhang, Jun Bi (2017) Development of land use regression models for $PM_{2.5}$, SO2, NO2 and O3 in Nanjing, China, *Environmental Research*,Volume 158, 542-552

[11] European Environment Agency. (2021, December 3). Air Quality Standards. https://www.eea.europa.eu/themes/air/air-quality-concentrations/air-quality-standards

[12] U.S. Environmental Protection Agency (2012). REVISED AIR QUALITY STANDARDS FOR PARTICLE POLLUTION AND UPDATES TO THE AIR QUALITY INDEX (AQI)

[13] X Li et al (2017) The Impact of Meteorological Factors on $PM_{2.5}$ Variations in Hong Kong *IOP Conf. Ser.: Earth and Environmental Science* 78 012003

[14] Amos P.K. Tai, Loretta J. Mickley, Daniel J. Jacob (2010) Correlations between fine particulate matter ($PM_{2.5}$) and meteorological variables in the United States: Implications for the sensitivity of $PM_{2.5}$ to climate change, *Atmospheric Environment*, Volume 44, Issue 32, 3976-3984,

[15] Karra, Kontgis, et al (2021) "Global land use/land cover with Sentinel-2 and deep learning." *IGARSS 2021-2021 IEEE International Geoscience and Remote Sensing Symposium. IEEE*

[16] Chen Z-Y, Jin J-Q, Zhang R, Zhang T-H, Chen J-J, Yang J, Ou C-Q, Guo Y. (2020) Comparison of Different Missing-Imputation Methods for MAIAC (Multiangle Implementation of Atmospheric Correction) AOD in Estimating Daily $PM_{2.5}$ Levels. *Remote Sensing*; 12(18):3008

[17] Witthuhn, J., Hünerbein, A., and Deneke, H. Evaluation of satellite-based aerosol datasets and the CAMS reanalysis over the ocean utilizing shipborne reference observations *Atmos. Meas. Tech.* 13, 1387–1412

[18] Schneider, R., Vicedo-Cabrera, A. M., Sera, F., Masselot, P., Stafoggia, M., de Hoogh, K., Kloog, I., Reis, S., Vieno, M., & Gasparrini, A. (2020). A Satellite-Based Spatio-Temporal Machine Learning Model to Reconstruct Daily $PM_{2.5}$ Concentrations across Great Britain. *Remote sensing*, 12(22), 3803.

[19] Maity, A., & Sherman, M. (2012) Testing for Spatial Isotropy Under General Designs. *Journal of statistical planning and inference* 142(5), 1081–1091 https://doi.org/10.1016/j.jspi.2011.11.013

=

[20] Jorge Mateu, Emilio Porcu, Pablo Gregori (2008) Recent advances to model anisotropic space-time data. *Stat. Methods Appl.* 17(2): 209-223

[21] Stephen S Lim et al. (2012) A comparative risk assessment of burden of disease and injury attributable to 67 risk factors and risk factor clusters in 21 regions, 1990–2010: a systematic analysis for the Global Burden of Disease Study 2010, *The Lancet*, Volume 380, Issue 9859, Pages 2224-2260,

[22] J.H.S. de Baar, R.P. Dwight, H. Bijl, (2013) Speeding up Kriging through fast estimation of the hyperparameters in the frequency-domain *Computers Geosciences*, Volume 54, Pages 99-106