Spatiotemporal forecast of NO2 in Madrid, Spain

Meng Lu

October 2017

1 Introduction

The number of air quality measurement stations is often limited to quantify air pollution continuously in space and time. Other data sources such as physical and chemical model simulations and spatiotemporal correlations may play an important role in improving air pollution interpolation. In this study, we investigated different Spatiotemporal and non-spatiotemporal statistical models and integrated historical records and predictions from a MACC physi-chemical model simulations to station measurements to predict air quality pollution.

We compared the results of air prediction from linear and non-linear regression models, geostatistical models, and random forest models. The linear and non-linear regression models assume no spatial correlation and purely predict using the MACC model simulations. We consider an ordinary kriging and an universal Kriging model to integrate spatial correlation and predict air pollution. For each linear/nonlinear regression, geostatistical, and random forest models, we attempted different methods and attempted to achieve the best results from these models.

The objectives are to find the best prediction method with available data. The objective is guided by three research questions:

- Can spatial correlation between a limited number of air quality measurement stations be used to improve air quality prediction?
- Can the MACC model simulation improve air quality prediction?
- How can we construct the random forest model to predict air quality with station measurements, MACC simulation, and temporal variables? Does the inclusion of time series characters, such as trend and harmonic terms improve model prediction?

2 Data

Three years of hourly NO2 station measurements and MACC simulations are available. There are in total 24 stations distributed in Madrid, Spain. These

station measurements are used to to train models and to validate the modeling results.

The distribution of MACC forecast of NO2 shows very close to Gaussian. The distribution of the measured NO2 is a skewed Gaussian distribution with mean around 33.

3 Method

We consider a variety of linear, non-linear regression models, Ordinary and Universal kriging methods, and random forest methods to predict NO2 spatiotemporally. Each of these methods are described in detail.

The NO2 predicting methods considered and compared are:

1. Ordinary kriging of MACC simulation
2. Linear regression model with MACC, harmonic terms and time of day (TOD) as potential independent variables.
3. Polynomial and GAM (general additive model) with MACC as independent variables.
4. Universal Kriging with MACC simulation.
5. Random forest prediction with MACC and different temporal variables.

3.1 Interpolation of MACC simulation

We then computed a pooled variogram from MACC forecasts of 200 sampled time stamps, and fitted the variogram with an Ste model. This variogram model is then used to interpolate MACC forecasts to the city grid and to each of the measuring stations.

3.2 Linear regression model with MACC as an independent variable.

The models in this section investigate linear or non-linear relationship between MACC simulation and station measurements. Linear, polynomial, and general additive models are attempted. (Adjusted) R square and model complexity were considered in model selection.

Five models are attempted to investigate the relationship between station measurements and spatially interpolated MACC forecast. Model 2a and 2b are linear regression models. In model 2b, TOD (time of day) and harmonic terms are used as additional independent variables. The models are built with station measurements and corresponding MACC simulations from 2012-01-01 to 2015-12-31 (table 1).

The harmonic terms $|sin(2\pi wt)+\phi|$, with $w = 1/(356*2*2)$ are used to fit the half year positive cyclical pattern. 45% percent of variance could be explained by this model ($lm(stationmeasurements \sim MACC + |sin(a)| + |cos(a)|)$).

Model	Method	Independent variable
2a	linear regression	MACC
2b	linear regression	MACC, TOD, Harmonics
2c1	second order polynomial	MACC
2c2	third order polynomial	MACC
2d	general additive model	MACC

Table 1: Linear and non-linear regression models and the independent variables of the model. TOD: time of day. Harmonics: $|sin(a)| + |cos(a)|$

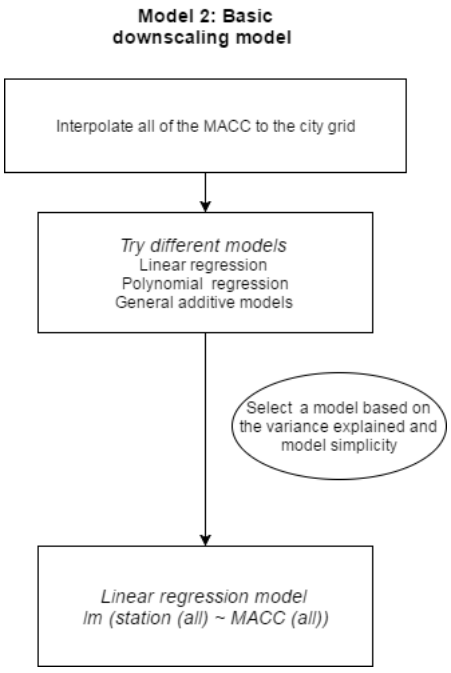

Figure 1: The workflow of model 2. The model 2a, which is a linear regression model with MACC as independent variable has recieved the best results and used to compare with models using other methods.

3.3 Universal Kriging

The basic downscaling models do not utilise spatial information. We attempted with Universal Kriging to utilise spatial correlations. Two models are attempted: 1) with residual variograms with MACC simulations as a regressor, and 2) with residual variograms with MACC simulations and locations as regressors.

We sampled 300 time stamps, compute a pooled variogram (average over the semivariance of each time stamp).

3.3.1 Manually fitting a variogram model to the variogram

As no spatial correlation is revealed in the variogram, fitting a variogram model automatically to it is not useful. We define the sill and nugget of a variogram model to fit to the variogram, then apply universal kriging using this variogram model to interpolate station measurements to the city grid and forecast to the next halfday.

3.3.2 Universal kriging models

fig. 2 shows the workflow the model 3a-c. The first two models (3a and 3b) use spatial data from the most recent time stamp to fit the data. The drawback of these two models is that the measurements or MACC forecasts at one or more stations may strongly affect the linear regression model (i.e. the relationship between MACC forecasts and station measurements may be inverted). These models 3a and 3b are more suitable to the situation that more stations (e.g. more than 100) are available or when there is a lack of historical time series data. In this study, we used data from all the locations and times to fit the linear regression model, and applied Kriging to the linear regression model residuals.

3.3.3 residual variogram

The residual variogram of a linear regression model with MACC simulations as a regressor. It could be observed that the spatial correlation is not revealed from the 24 station measurements. Three possible reasons are: 1) there are too few observations (stations), 2) there is a lack of information from short-distance placed stations, 3) due to the design of the air pollution measuring station network, the stations are placed where the pollution is suspected to be high, e.g., near factories, in a canyon. distance semivariance

3.3.4 pooled variogram

Pooled variogram is computed from spatiotemporal points, which uses more information than only using data from a time stamp and reduces the influence of extreme values. Firstly we sample 300 time stamps that contain no missing data. Then, we compute a pooled residual variogram (average semivariance) of (1) a linear regression model with MACC simulations as a regressor, and (2) a linear regression model with macc simulations as a regressor.

The pooled residual variogram with MACC as independent variables (figure 5) contains a little more information than the variogram of a time stamp. Spatial correlation is not shown in the pooled variogram. The pooled residual variogram (figure 6) with MACC and locations as independent variables also does not reveal spatial correlation.

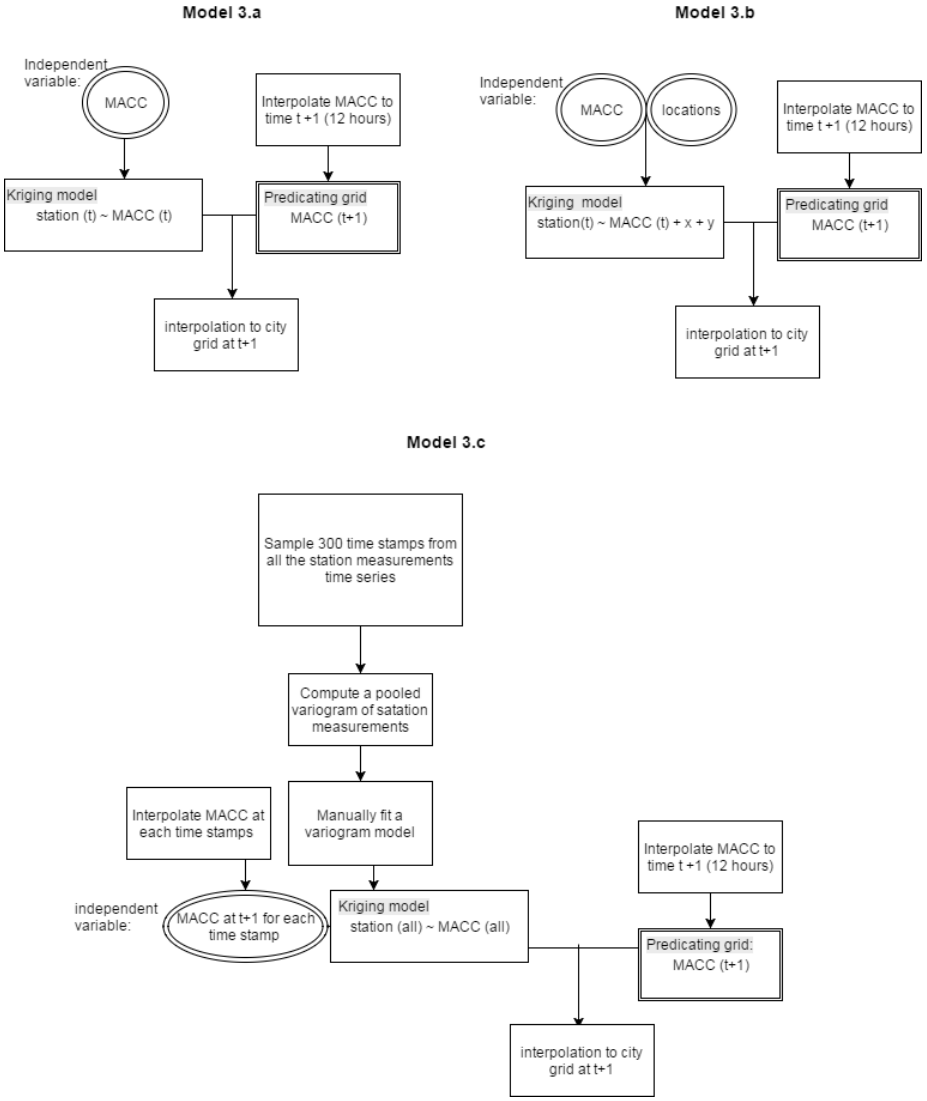

Figure 2: workflow the universal krigging models. Model 3a: using station measurements and MACC simulations at the same time to fit a model, then using MACC simulations at time t+1 to predict. Model 3b: Using station measurements and MACC simulations at the same time, as well as locaitons to fit a model, then using MACC simulations at time t+1 as well as locations to predict. Model 3c: Using all the measurements (about 4 years of halfday data) to fit a linear regression model, then apply universal Kriging with fixed coefficients from the linear regression model. The model is applied to the MACC simulations at time t+1 to predict air quality.

3.4 Random forest

This study focuses on using random forest regression to forecast No2 MACC simulation, day of year (DOY), time of day (TOD), lagged MACC simulations and harmonic terms are used as independent variables.

The first half of the data (i.e. the first 2 years of time series, in total 1420 observations) are used to train the forest. The second half of the data (i.e. the last 2 years of time series, in total 1428 observations) are used to test the model. The missing values are filled using spline interpolation.

The permutation importance (Altmann et al., 2010) is used as an important measure. It is reported in Altmann et al. (2010) that this measure corrects the bias of other conventional importance measures (i.e., Gini importance and vairable importance) for categorical and grouped variables.

The R package "ranger" is used to perform random forest. The "ranger" package aims at optimising for multi-dimensional anaylsis. In this study, it is found that ranger is faster than the "randomforest package". I did not compare all the results from difference R packages for random forest, it might be of interest to do so.

Table 1 shows the variables that are used in each model. The harmonic term that is used in model 2 ($|sin(2\pi wt) + \phi|$) is used in the one of the random forest model. The differences between model 4.2a and model 4.2b is that the model 4.2b treats DOY as factors while model 4.2a treats it as numbers. As using a lagged MACC (as is used for Ozone) to forecast Ozone obtains the lowest accuracy, and is slow, for NO2 this model is taken off.

Model	Independent variable
4.1	MACC simulations, DOY, harmonic term, TOD
4.2a	MACC simulations, DOY, TOD (factors)
4.2b	MACC simulations, DOY (factors), TOD (factors)
4.3	DOY, MACC simulations
4.4	MACC simulations

Table 2: Independent variables used in each model. DOY: day of year, TOD: time of day

3.4.1 Accuracy assessment

Obtaining the importance of each variable using the permutation importance proposed by Altmann (2010). Using half of the time series to train the model and another half to test the model. The models are assesssed using out of bag (OOB) prediction error and OOB R square (internal cross validation) of the trainning set. Same as the accuracy assessment of the model 1-3, we assess the accuracy using median, IQR (interquartile range), Bias(the mean of errors), RMSE (Root Mean Square Error), and MAE(Mean Absolute Error).

All the models were firstly trained and tested with a single station measuremntes and then trained and tested with all the station measurements.

4 Results

Errors (differences between predictions and station measurements) of applying the developed models (the best model of each report) to the testing set (the second half of the all the time series). Random forest models obtained lower errors in terms of mean and median. Model 4.3 obtained the best accuracy, and it is simple.

Model	Method	Meidan	IQR	Mean	RMSE	MAE
1	OK	-18.29	22.74	-22.20	28.76	21.13
2a	LM	4.3	23.2	5.32	19.73	16.00
2b	LM	-2.51	26.68	-5.90	23.09	17.44
3c	UK	-6.26	35.2	-0.75	31.10	24.28
4.1	RF	1.90	19.71	-0.26	19.54	14.80
4.2a	RF	1.11	20.97	-2.69	20.13	15.21
4.2b	RF	1.17	21.00	-2.66	20.16	15.24
4.3	RF	1.10	21.60	-0.60	20.58	15.76
4.4	RF	0.64	23.59	-1.15	22.21	16.92

Table 3: Accuracy assessment results of the best models of method OK (ordinary kriging), LM (linear regression) and UK (universal kriging), and all the models of the RF (random forest) regression. MACC: MACC simulations. TOD: time of day. DOY: day of year. har: harmonic terms.

4.1 Model 1

Interpolation of MACC forecasts: Strong spatial correlation are shown between MACC forecasts for all the parameters. Relatively low interpolation error.

4.2 Model 2a-d

Basic downscaling model: There is a linear relationship between MACC forecasts and station measurements of Ozone, the polynomial and general additive model were also applied but a linear model is selected for Ozone. The linear relationship between NO2 station measurements and MACC simulations are less obvious comparing to the Ozone case. For different models, Polynomial model of different orders and a general additive model were attempted. Harmonic terms were also used to fit the model.

4.3 Model 3a-c

Universal kriging: There are 14 stations available for Ozone and 24 stations available for NO2. Possibly because of this, Applying universal kriging to NO2 has received better results comparing to Ozone. Pooled residual variograms were used and calculated. No spatial correlation between stations for all the variables

can be indicated from the variograms of stations. So a variogram model is manually decided (by assigning a variogram sill and a nugget) by observing the pooled variogram, assuming spatial correlation exists within shorter distances. There are spatial trends for all the variables, therefore locations are used as regressors together witht the MACC simulation.

4.4 Model 4.1-4-6

Random forest regression: One or more of the variables of MACC simulation, day of year (DOY), time of day, lagged MACC simulations and harmonic terms are used as independent variables. Random forest methods have obtained improved results comparing to other methods for both Ozone and NO2. For both of the methods, using MACC and DOY as independent variables received the best results.

In summary, for both Ozone and NO2, the random forest regression models can obtain the lowest RMSE (the best accuracy and lowest variance). Spatial correlation may be used in random forest (as distance) to further improve the accuracy. If random forest is not available, the linear regeression model 2a (for Ozone) or 2b (for NO2) is very simple and the result it obtained is not bad. Model 2a (for Ozone) or 2b (for NO3) obtains a bigger variance and higher MAE, but is significantly improved from model 1. The error of Model3c is accessed different from other models, as spatial errors are assessed. As we have mentioned in report 3, a variogram model which assumes spatial correlation within distances smaller than between stations is manually fitted, it is possible that model 3c has some advantage when applying the model to predict the air quality of the whole city grid. Among all the random forest models, model 4.3 could be the most favorable model to use due to its simplicity and high accuracy. Apart from model 1, most methods (universal kriging model, linear regression model and Random forest) obtained better results for NO2 than for Ozone.