Prediction of PM10 and PM2.5 concentration using land use data and spatial correlation

Tomoshige Nakamura[*] Mihoko Minami [†]

Abstract

We are concerned with the effect of a particulate matter on our health in Japan. When we estimate the effect of a particulate matter on our health, we would like to have observations of particulate matter concentrations in study areas. However, particulate matter concentrations are not observed in a part of cohort study areas, so we need to estimate PM concentrations in unobserved areas to estimate the effect of particulate matter on our health. In such a case, it is common to use some estimation method (e.g. Inverse distance weighting, land use regression) to obtain estimates of particulate matter concentrations at unobserved areas, then replace missing values with these estimates. However, the variance of the effect of particulate matters on our health may be underestimated using such methods.

This paper discusses the problem when we substitute estimates as observed values to analyze the effect of particulate matters on our health through simulation, and construct the model for PM10 spatial distribution in Japan to use a Bayesian approach to avoid previously mentioned problems.

Key Words: Particulate matter, Regression imputation, Bayesian approach, Spatial model

1. Introduction

Air Pollution by particulate matter such as PM2.5 or PM10 is an issue that attracts increasing public concern. Numerous epidemiological studies show the association between above-mentioned air pollutant and short-term mortality and long-term mortality(see WHO 2005), and there are several studies about the modeling relationship between PM2.5 and respiratory organs, and mortality, (see Fuentes et.al 2006, McBrige et.al 2007). We would like to investigate the association between particulate matter and various diseases using long-term health survey data on health hazard evaluation area in Japan.

ESCAPE Study (see Eeftens et.al 2012, Stafoggia et.al 2014) investigates long-term effects on human health of exposure to air pollution in Europe. In ESCAPE study, there are study areas which particulate matter concentrations are not observed, so they constructed the model for particulate matters using land use data observed in areas where particulate matters were observed, then they fitted the model to areas which particulate matters are not observed, calculated fitted values (e.g. Regression estimates), and substituted regression estimates as the observed values in that area.

In ESPAPE study, there are only a few study areas which are not observed particulate matter concentrations, but in our study in Japan, at more than 70% of study areas, particulate matter concentrations are not observed. If we use the same method in Japan, our estimates and inference about the relationship between particulate matter and our health may be biased.

In this paper, firstly we point out the issue of substituting regression estimates as the observed values in analyzing particulate matter effects on our health by simulation, then describe the other approach (i.e. Bayesian approach) and its problem in practice, and point

[*] Graduate School of Science and Technology, Keio University, 3-14-1 Hiyoshi, Kohoku-ku, Yokohama, Kanagawa 223-8522, Japan

[†] Department of Mathematics, Keio University, 3-14-1 Hiyoshi, Kohoku-ku, Yokohama, Kanagawa 223-8522, Japan

out both problems are caused by poor model for particulate matter concentrations. At last, we try to find what component is important to construct the model for particulate matter concentrations.

2. Simulation of Poisson regression with covariate all missing.

In this simulation, we suppose particulate matter concentrations are not observed in cohort study areas. Under this condition, we suppose to estimate the effect of particulate matter on patients number and event number observed in cohort study areas.

Let i be the index of cohort study areas, and j be the index of monitoring stations around the cohort study areas, where $i \in \{1, 2, \cdots, n\}$ and $j \in \{1, 2, \cdots, m\}$. Let Y_i be the number of patients or events in a cohort study area, X_i and X_j^* be a concentrations of particulate matter observed in a cohort study area and at a monitoring station, and Z_i and Z_j^* be a covariates for particulate matter concentrations in a cohort study area and at a monitoring station.

In this simulation, we assume particulate matter concentrations in cohort study areas are not observed, so X_i are all missing. (i.e. In cohort study area, Y_i and Z_i are observed, but X_i is missing. At monitoring stations, X_j^* and Z_j^* are observed). To estimate the effect of Y on X, we construct the model for X using Z observed at monitoring stations, then we use the constructed model for X to estimate the missing X at cohort study areas. At last we analyze the effect of Y on X. Next, we describe the model to generate the hypothetical data.

2.1 Models used to generate hypothetical data

We assume Y_i is distributed as poisson distribution.

$$Y_i \overset{iid}{\sim} \mathrm{Pois}(\lambda_i) \text{ where } \log \lambda_i = X_i' \beta \tag{1}$$

where $X_i' = (1, X_i)$. In practice, X_i' include other covariate which relate to Y_i, but in this simulation, for simplicity, we assume X_i' is two dimensional matrix. We suppose X_i is distibuted as normal.

$$X_i = Z_i \alpha + e_i \text{ where } e_i \overset{iid}{\sim} \mathrm{N}(0, \tau^2) \tag{2}$$

where Z_i is covariate for particulate matter concentration, for example, temperature, humidity, and other meteorological covariates, elevation, traffic amounts, and other land use data. At last, we set sample sizes and parameters as follows.

- $\boldsymbol{\beta} = (\beta_0, \beta_1)' = (1, 0.5)'$
- $\boldsymbol{\alpha} = (-0.7, 0.5, 1.2)'$
- $\tau = 3.5$
- $n = m = 200$

2.2 Models used to analyze hypothetical data

To analyze generated hypothetical data, we fit following two models and estimate MLE $\hat{\boldsymbol{\beta}}$ and its confidential interval $\mathrm{CI}(\boldsymbol{\beta})$. Model - I we suppose the X_i are observed and using X_i as covariates for Y_i, and model - II we use the expectation of X_i as covariates for Y_i.

$$\begin{array}{rl} \text{model - I} & Y_i \overset{iid}{\sim} \mathrm{Pois}(\lambda_i) \;\; \lambda_i = X\beta \\ \text{model - II} & Y_i \overset{iid}{\sim} \mathrm{Pois}(\lambda_i) \;\; \lambda_i = \mathrm{E}[X]\beta \end{array}$$

Model - ii is corresponding to the analysis of the effect of particulate matter concentrations on our health when particulate matter concentrations in cohort study areas are missing.

In practice, $\mathrm{E}[X_i]$ cannot be observed, so we estimate $\hat{\alpha}$ using monitoring stations data and replace $\mathrm{E}[X] = Z\boldsymbol{\alpha}$ as $Z\hat{\boldsymbol{\alpha}}$. In this simulation, we assume we know the true value of α.

2.3 Simulation results

In this simulation, we iterate 2000 times hypothetical data generation and analyzing its hypothetical dataset, and computed $\hat{\boldsymbol{\beta}}$ and $\mathrm{CI}(\boldsymbol{\beta})$. We sort two thousands of confidence intervals $\mathrm{CI}(\boldsymbol{\beta})$ in ascending order of $\hat{\beta}_1$ values. Left panel of Figure1 depicts confidential intervals of β for model -I, and right panel for model II. Left panel shows 96% of confidence intervals contain the true value $\beta_1 = 0.5$, right panel shows 15% of confidence intervals contain the true value. So, analysis under the model I is verified theoretically, but analysis under the Model II, substituting regression estimates as observed values, almost all of confidential interval don't contain true value, and length of estimated confidence interval is not proper for inference.

A magnitude of variance parameter of X, τ, control the results of simulation. If τ become large, less confidence intervals contain the true value. If τ become small, more confindence intervals contain the true value. In practice, we replace τ with estimate $\hat{\tau}$. If constructed model for particulate matter is poor, estimate of τ become large, and then the confindence interval for $\hat{\beta}$ become less reliable. Moreover, now we use the true value for α, but in practice, we estimate it, so, result may become worse.

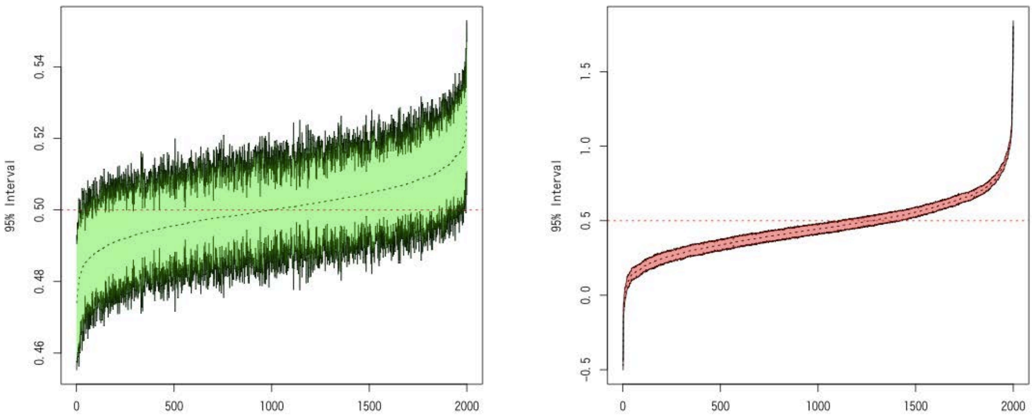

Figure 1: Plot of 2000 confidence intervals of β_1 obtained by simulations. Left panel shows confidence intervals under Model I, and right panel shows confidence intervals under Model II. Confidence intervals are reordered in ascending order by $\hat{\beta}_1$ values. "- - -" shows $\hat{\beta}_1$ values.

As we show the above, substituting regression estimates for missing X_i may lead wrong inference. This problem can be avoided only when τ is small enough, that is, when we can construct the accurate model for X. However, if we don't have enough data for constructing the model for X, then the approach, substituting the regression estimates as observed values, is not appropriate. In missing data analysis, Bayesian approach is often used for this kind of problems. When we analyze missing data by Bayesian approach, confidence interval for parameters can be appropriately obtained under suitable condition. However Bayesian approach also has problems. If we can't contruct the model each layer of hierarchy of bayesian model properly, the length of confidence interval become too large and we cannot lead significant result. As a result, both methods demand an accuracy model for particulate matter concentrations. Next section, we try to reveal what components (such as spatial structure, landuse data, meteorological data) are important to construct the model for particulate matter concentrations in Japan.

3. Prediction of particulate matter concentration in Japan

In this section, we consider the component which we should contain in model for particulate matter prediction. Firstly, we show the present situation for particulate matter observation. Secondly, describe about data we use. Thirdly, we fit several models and check the result.

3.1 Monitoring stations around cohort study area

We plotted monitoring stations around the cohort study areas in Figure.3.1. There are 264 monitoring stations around the cohort study area, and 240 of them observe PM10 concentrations, but for PM2.5 there are only 30 monitoring stations. So there are not enough data for PM2.5 to model PM2.5 concentrations. In the following section, we target to construct the model for PM10 annual concentrations.

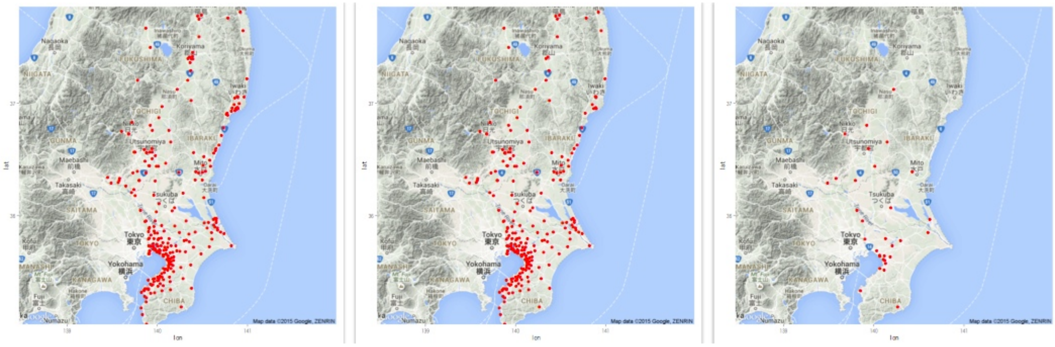

Figure 2: Red dots show the station locations in 2011 and 2012. Left panel:all monitoring stations around study areas. Middle panel : monitoring stations observe PM10 concentrations. Right panel : monitoring stations observe PM2.5 concentrations.

3.2 Data Description

Air pollution and meteorological data at each monitoring stations can be downloaded from National Institute for Environmental Studies, Japan (http://www.nies.go.jp/). We extracted the data from April in 2011 to March 2012 in 6 prefectures around the cohort study areas, and variables observed at each monitoring stations are different. Each variables in dataset are observed hourly, so we aggregated the data and calculated daily average, and then calculated annual average. sample size is 229, we selected following variables as covariates for PM10 annual concentration.

land use data	meteorological data
Longitude	annual average of NO_2
Latitude	annual average of NO
Elevation	annual average of Wind speed
Prefecture	annual average of temperature
Area application around monitoring station	
NOxPM combating Area or Not	

Table 1: Covariates for PM10 annual concentration. we use two types of covariates. Land use data is GIS data obtained at monitoring stations, and meteorogical data is observed at monitoring stations.

3.3 Model Candidate and the results

Our target is clarify the component which play a important role in predicting the particulate matter concentrations in Japan. To do this, we consider the several model candidates, and compare fitted result. Now, we denote $\boldsymbol{s}_i$ as monitoring station location, and Y_i as PM10 concentration observed at location $\boldsymbol{s}_i$, and $\boldsymbol{Y} = (Y_1, \cdots, Y_n)^T$. We suppose a following model assumption for all candidate models.

$$\boldsymbol{Y} \sim \mathrm{N}(X\boldsymbol{\beta}, \Sigma) \tag{3}$$

where Σ is covariance matrix with spatial structure for PM10 concentration, X is the covariates for $\boldsymbol{Y}$. If we don't include spatial correlation in the model, we assume $\Sigma = \tau^2 I$, and when we include it, we assume following structure.

$$\Sigma = \tau^2 I + \sigma^2 H(\phi) \quad \text{where} \quad H(\phi)_{ij} = \exp(-\phi\|\boldsymbol{s}_i - \boldsymbol{s}_j\|^2) \tag{4}$$

There are many candidate for structure of Σ, but in this analysis we use the above. We summarize models for comparing in following Table3.3.

	Covariantes in X	Spatial Structure for Σ
model i	only intercept	Yes
model ii	land use data and meteorological data	No
model iii	meteorological data	Yes
model iv	land use data and meteorological data	Yes

Table 2: First column shows covariates including each models. description of covariates denoted in 3.2. Second column shows Σ include a spatial structure or not in each models.

In order to compare the candidate models, we compute three fold cross-validation estimates of R^2 and mean squared residual. To compute these estimates, we sample from posterior of parameters $(\boldsymbol{\beta}, \tau, \sigma, \phi)$ by MCMC, and compute $\mathrm{E}[\boldsymbol{y}_0|\boldsymbol{Y}]$ using MCMC samples (where $\boldsymbol{y}_0$ is concentration of PM10 at unknown place $\boldsymbol{s}_0$).

The result of estimated R^2 and mean residuals sum of squares under candidate models are summirized in following table. From this table, model ii, that does not contains spatial structure, is poor, compared to that of othe models. Without model ii, we can see progressice improvement in prediction capability, moving from model i to model iv. These result means that spatial structure, meteorological covariantes and land use covariates are all play a important role to predict PM10 concentration in Japan. But the fact $R^2 = 0.40$ represent the accuracy of these models is not still good.

model	i	ii	iii	iv
RMSE	10.66	14.31	10.40	9.31
R^2	0.32	0.19	0.37	0.41

4. Discussion

This article presents, to impute regression estimates as observed value may cause wrong inference when we use Poisson regression model to estimate the effect of particulate matters on our health. So we suppose to use Bayesian approach to avoid this problem. However there is another problem to be solved. If we cannot construct the model for particulate matter concentrations, then confidential interval of parameters become large and we cannot lead meaningful result. Therefore we try to construct the model for PM10 concentrations and find what component plays an important role to predict PM10 concentrations. As a result, We find spatial structure, land use variables and meteorological variables are all important component to construct model in Japan, but prediction accuracy is still not good.

There are two future works we have. One is about simulation of Poisson regression model. The problem caused by substituting regression estimate as observed values in Poisson regression model is only derived by simulation, not by theoretically. We try to reveal this result from theoretical aspect. Anothe one is about models for PM10. In previous section, we construct a model for PM10 with limiteed variavles as covariates. To construct a model for PM10 accurately, we try to find out crucial covariates for PM10.

REFERENCES

World Health Organization(2006) “Air quality guidelines -Global update 2005- Particulate matter, ozone, nitrogen dioxide and sulfur dioxide,” pp. 217-280.

Mcbride, S. J., R. W. Williams, and J. P. Creason. (2007) “Bayesian hierarchical modeling of personal exposure to particulate matter,”. *Atmospheric Emvironment. Elsevier Science Ltd, New York, NY*, 41(29):6143-6155.

Fuentes, M., Song, H.-R., Ghosh, S. K., Holland, D. M., andDavis, J. M. (2006) “Spatial association between speciatedne particles and mortality”. *Biometrics* 62, 855863.

Marloes Eeftens, Rob Beelen, Kees de Hoogh, Tom Bellander, Giulia Cesaroni, Marta Cirach, Christophe Declercq, Audrius Ddel, Evi Dons, Audrey de Nazelle, Konstantina Dimakopoulou, Kirsten Eriksen, Grgoire Falq, Paul Fischer, Claudia Galassi, Regina Grauleviien, Joachim Heinrich, Barbara Hoffmann, Michael Jerrett, Dirk Keidel, Michal Korek, Timo Lanki, Sarah Lindley, Christian Madsen, Anna Mlter, Gizella Ndor, Mark Nieuwenhuijsen, Michael Nonnemacher, Xanthi Pedeli, Ole Raaschou-Nielsen, Evridiki Patelarou, Ulrich Quass, Andrea Ranzi, Christian Schindler, Morgane Stempfelet, Euripides Stephanou, Dorothea Sugiri, Ming-Yi Tsai, Tarja Yli-Tuomi, Mihly J Varr, Danielle Vienneau, Stephanie von Klot, Kathrin Wolf, Bert Brunekreef, and Gerard Hoek. (2012) “Development of Land Use Regression Models for PM2.5, PM2.5 Absorbance, PM10 and PMcoarse in 20 European Study Areas; Results of the ESCAPE Project,” *Environmental Science and Technology*, 46 (20), pp. 11195-11205

Stafoggia M, Cesaroni G, Peters A, Andersen ZJ, Badaloni C, Beelen R, Caracciolo B, Cyrys J, de Faire U, de Hoogh K, Eriksen KT, Fratiglioni L, Galassi C, Gigante B, Havulinna AS, Hennig F, Hilding A, Hoek G, Hoffmann B, Houthuijs D, Korek M, Lanki T, Leander K, Magnusson PK, Meisinger C, Migliore E, Overvad K, stenson CG, Pedersen NL, Pekkanen J, Penell J, Pershagen G, Pundt N, Pyko A, Raaschou-Nielsen O, Ranzi A, Ricceri F, Sacerdote C, Swart WJ, Turunen AW, Vineis P, Weimar C, Weinmayr G, Wolf K, Brunekreef B, Forastiere F. (2014). “Long-term exposure to ambient air pollution and incidence of cerebrovascular events: results from 11 European cohorts within the ESCAPE project,” *Environ Health Perspect* 122, pp. 919925