Use of data imputation tools to reconstruct incomplete air quality datasets

Quinteros, María Elisa; Lu, Siyao; Blazquez, Carola; Cárdenas-R, Juan Pablo; Ossa, Ximena; Delgado-Saborit, Juana María; Harrison, Roy M.; Ruiz-Rudolph, Pablo

DOI:
10.1016/j.atmosenv.2018.11.053

Citation for published version (Harvard):
Quinteros, ME, Lu, S, Blazquez, C, Cárdenas-R, JP, Ossa, X, Delgado-Saborit, JM, Harrison, RM & Ruiz-Rudolph, P 2019, 'Use of data imputation tools to reconstruct incomplete air quality datasets: a case-study in Temuco, Chile', *Atmospheric Environment*, vol. 200, pp. 40-49. https://doi.org/10.1016/j.atmosenv.2018.11.053

A paper to be submitted to *Atmospheric Environment*

Use of data imputation tools to reconstruct incomplete air quality datasets: A case-study in Temuco, Chile

María Elisa Quinteros[a, b], Siyao Lu[c], Carola Blazquez[d], Juan Pablo Cárdenas-R[e], Ximena Ossa[f], Juana-María Delgado-Saborit[g,h,i, j], Roy M. Harrison[g,k], Pablo Ruiz-Rudolph[l*]

[a] Programa Doctorado en Salud Pública, Instituto de Salud Poblacional, Facultad de Medicina, Universidad de Chile, Independencia 939, Independencia, Santiago, Chile.

[b] Departamento de Salud Pública. Facultad de Ciencias de la Salud, Universidad de Talca, Avenida Lircay s/n, Talca, Chile.

[c] Department of Environmental Health Sciences, University of Michigan, 1415 Washington Heights, Ann Arbor, MI 48109, EE. UU.

[d] Department of Engineering Sciences, Universidad Andres Bello, Quillota 980, Viña del Mar, 2531015, Chile.

[e] Departamento de Ingeniería en Obras Civiles. Instituto del Medio Ambiente, Universidad de La Frontera, Avenida Francisco Salazar 01145, Casilla 54-D, Temuco, Chile.

[f] Departamento de Salud Pública y Centro de Excelencia CIGES, Universidad de la Frontera, Caro Solar 115, Temuco, Chile.

[g] Division of Environmental Health and Risk Management, School of Geography, Earth and Environmental Sciences, University of Birmingham, Edgbaston Birmingham B152TT, UK.

[h] ISGlobal Barcelona Institute for Global Health, Barcelona Biomedical Research Park, Doctor Aiguader 88, 08003, Barcelona, Spain.

i Pompeu Fabra University, Plaça de la Mercè 10, 08002, Barcelona, Spain.

[j] Spanish Consortium for Research on Epidemiology and Public Health (CIBERESP), Instituto de Salud Carlos III, Avenida Monforte de Lemos 5, E-28029, Madrid, Spain.

[k] Department of Environmental Sciences / Center of Excellence in Environmental Studies, King Abdulaziz University, PO Box 80203,Jeddah, 21589, Saudi Arabia.

[l] Programa de Salud Ambiental, Instituto de Salud Poblacional, Facultad de Medicina, Universidad de Chile, Independencia 939, Independencia, Santiago, Chile.

[*]Corresponding author

Corresponding Author

Pablo Ruiz-Rudolph. Programa de Salud Ambiental, Instituto de Salud Poblacional, Facultad de Medicina, Universidad de Chile, Independencia 939, Independencia, Santiago, Chile; pabloruizr@uchile.cl; phone (+56-22-978-6379)

List of tables

List of figures

List of supplemental tables

List of supplemental figures

Abstract

Missing data from air quality datasets is a common problem, but it is much more severe in small cities or localities. This poses a great challenge for environmental epidemiology as high exposures to pollutants worldwide occur in these settings and gaps in datasets hinder health studies that could later inform local and international policies. Here, we propose the use of imputation methods as a tool to reconstruct air quality datasets and applied this approach to an air quality dataset in Temuco, a mid-size city in Chile as a case-study. We attempted to reconstruct the database comparing five approaches: mean imputation, conditional mean imputation, K-Nearest Neighbor imputation, multiple imputation and Bayesian Principal Component Analysis imputation. As a base for the imputation methods, linear regression models were fitted for $PM_{2.5}$ against other air quality and meteorological variables. Methods were challenged against validation sets where data was removed artificially. Imputation methods were able to reconstruct the dataset with good performance in terms of completeness, errors, and bias, even when challenged against the validations sets. The performance improved when including covariates from a second monitoring station in Temuco. K-Nearest Neighbor imputation showed slightly better performance than multiple imputation for error (25% vs. 27%) and bias (2.1% vs. 3.9%), but presented lower completeness (70% vs. 100%). In summary, our results show that the imputation methods can be to a certain extent successful in reconstructing air quality dataset in a real-life situation.

Keywords:

Wood-burning; Air pollution; Missing data; Multiple imputation; Environmental epidemiology; Single imputation.

1 Introduction

Missing data in environmental monitoring is a common problem worldwide, but can be much more severe in small cities or localities (Green and Sánchez, 2012). Some conditions that drive this higher than usual losses in air quality networks include lack of coverage and representativeness, main localization in capital cities, stations run manually, instrument failures, and human errors (Riojas-Rodriguez et al., 2016; Toro A. et al., 2015). This is a great challenge for environmental epidemiology, as higher exposures to pollutants often occur in these settings, particularly in lower income countries, and this lack of data could later hinders health impact assessments (Pascal et al., 2013) or epidemiological studies that in turn could inform local and international policies (World Health Organization, 2016).

Missing data is, at its root, a statistical problem. It represents a form of measurement error that may both bias the sample and decrease sample size (Little and Rubin, 1987). Proper handling of missing data should be observed in all statistical analyses, and the methods to be used depend on the missing mechanism (Little and Rubin, 1987). Basically, there are three possible mechanisms: i) missing completely at random (MCAR), where missing data are unrelated to either observed or unobserved data; ii) missing at random (MAR), where missing data are partially related to observed data; and, iii) missing not at random (MNAR), also known as non-ignorable or non-response, where missing observations are related to values of the unobserved data (Little and Rubin, 1987).

When faced with missing data, researchers often employ the complete case approach, also called list-wise deletion, where the analysis is performed after deleting all observations with any missing data (van Buuren, 2012). As a result, sample size and statistical power is reduced, and bias may be introduced if data are MNAR. Another common approach is single imputation, where missing data are replaced or imputed with a single value provided by a suitable method such as mean imputation, random imputation, or conditional mean imputation. However, these methods may generate biased and unsatisfactory results, as the imputation error is neglected, and thus underestimating standard errors (Greenland and Finkle, 1995).

Since the mid-eighties more sophisticated approaches have been introduced, including expectation maximization, weighted estimating equation methods, and particularly, K-Nearest Neighbor, multiple imputation and imputation using Bayesian principal component analysis. The nearest neighbor imputation draws imputed values from the closeness observation based on the absolute difference between the linear prediction for the missing value and that for the complete values (Dixon, 1979). Multiple imputation is based on Bayesian methods, and its main purpose is to properly reproduce the variance/covariance matrix had the data been complete, thus providing valid inference under MAR assumptions (Little and Rubin, 1987). It uses an iterative form of stochastic imputation, creating multiples copies of the database, where missing values are replaced by imputed values from a posterior predictive distribution using the partially observed data. Subsequently, every database is analyzed and results are combined, including standards errors. Therefore, data uncertainty is incorporated in the process (Little and Rubin, 1987; Rubin, 1987). The Bayesian principal component analysis imputation involves Bayesian estimation of missing values with the iterative expectation maximization algorithm. This analysis is based on three processes: principal component regression, Bayesian estimation, and an expectation–maximization (EM)-like repetitive algorithm (Bishop, 1999).

Despite the fact that imputations tools are available in many statistical packages, they are not often used very in epidemiological studies (Klebanoff and Cole, 2008; Sterne et al., 2009; Stuart et al., 2009). Moreover, in environmental epidemiology the most common approaches have been to ignore them (i.e., the complete case analysis), to replace missing data based on prior knowledge, or to use single imputation, for instance, from a multiple regression (Roda et al., 2014). Some studies have included multiple imputation applied to air quality datasets (Junger and de Leon, 2009, 2015; Junninen et al., 2004; Roda et al., 2014), but overall its application remains scarce with few tests of performance in real-life situations and providing little guidance with respect to the application in other settings.

Here, we propose to use imputation methods as a tool to reconstruct air quality datasets and applying them to an air quality dataset in Temuco, a mid-size city in Chile as a case-study. Temuco resembles the problems faced in many small-medium cities in the world, whose datasets may be fragmented. It also faces a major environmental health problem being heavily impacted by residential wood-burning, as many southern Chilean cities, highlighting the importance of having full data for epidemiological studies (Díaz-Robles et al., 2008; Gómez et al., 2014; Villalobos et al., 2017). In this study, we attempt to reconstruct the database comparing five approaches: mean imputation, conditional mean imputation, K-Nearest Neighbor imputation, multiple imputation and Bayesian Principal Component Analysis imputation. The overall approach considers i) developing a standard regression model of $PM_{2.5}$ using available predictors that could explain the air pollutant concentration in the case study (i.e. meteorological and co-pollutants), ii) based on the best models, applying the imputation methods to complete the datasets, iii) building validation datasets by artificially removing data, and iv) assessing the performance of the methods in reconstructing the removed data in the validation sets. The application of the best method is expected to be used in a real-life situation in Temuco by completing the $PM_{2.5}$ datasets required to build a land-use regression model, which will later be used to estimate exposures in a health study of wood-burning air pollution and pregnancy outcomes (Ruiz-Rudolph, 2014).

2 Methods

2.1 Study Area

Temuco is a mid-size city of 290,000 inhabitants located in the Araucanía Region, in southern Chile (longitude 39.7°E; latitude 73.0°S) in a valley crossed by the Cautín river and surrounded by hills, native forest, and agricultural fields (Minsal, 2016). The "Great Temuco" is a conurbation of two cities: Temuco, to the north, and Padre Las Casas, to the south across the river (Figure 1). Temuco, and the Araucanía region in general, present a population of medium to low socioeconomic status, which is reflected by the 22.9% of the households that are classified as poor, and by the only 8.2 years of schooling on average of the head of the household (Ministerio de Desarrollo Social, 2011). The city experiences a Mediterranean climate with oceanic influence (Csb), with average temperatures close to 12ºC, rainfall above 1,000 mm per year, and marked seasonal differences, with cold, humid winters, and low wind speeds associated with poor air pollution dispersion (Ministerio de Medio Ambiente, 2014).

The study area has some characteristics different from other many Chilean cities but similar to many in the south. For example, the industrial activity in the area is low with agriculture being the main economic activity (Minsal, 2016). Known air pollution sources include some stationary emissions such as industrial wood- and coal-fired boilers associated with the processed woods industry (Ministerio del Medio Ambiente, 2015), and a medium-sized fleet of 67,800 motorized vehicles (INE, 2017). However, the largest aggregated source of $PM_{2.5}$ and PM_{10} is the residential wood-burning that is used throughout the city in winter for heating and cooking. More than 88% of homes have wood-stoves, and approximately 654,000 m^3 of wood are used per year (Gómez et al., 2014; Ministerio del Medio Ambiente, 2015; Molina Sepúlveda and Oyarzo Gómez, 2013; Villalobos et al., 2017).

2.2 Data sources

The Great Temuco has an air pollution monitoring network that measures PM_{10}, $PM_{2.5}$, SO_2, NO_x, O_3, CO, and meteorological variables. This network is run by the Ministry of the Environmental, and hourly data is available online (Ministerio de Medio Ambiente, 2017). The network is comprised by two stations in Temuco (*Las Encinas* and *Museo Ferroviario* stations) and another one in Padre Las Casas comprise the network (Figure 1). The three stations began $PM_{2.5}$ measurements in 2009. Since *Las Encinas* contains more the complete sets, we focus in reconstructing its full series of $PM_{2.5}$ from 2009 to 2014, so it can be later used to estimate historical exposures. Note that there is no available dataset capturing the regional contribution of air pollutants levels in the studied area. Additional meteorological data were obtained from the *Maquehue* station run by the Meteorological Office of Chile (Dirección Metereológica de Chile, 2016), which is located outside the urban area, about 3 kilometers south of the downtown area close to a former aerodrome.

2.3 Statistical analysis and imputation methods

Hourly air pollution data was converted to daily means according to the national legislation (Ministerio del Medio Ambiente, 2018). After an initial analysis of completeness, the missing data mechanism was diagnosed using two tests: Little´s MCAR test (Little, 1988) and the test of missingness (Schafer and Graham, 2002). The data distribution was explored for all variables through histograms, Q-Q plots and the Shapiro-Wilk to test normality (Figure S 1). As distributions of $PM_{2.5}$ and PM_{10} were heavily skewed, they were log-transformed, which improved their performance and were used in further analysis. Descriptive analyses were performed for all variables including mean, median, percentiles and measures of dispersion (Table S1- S 2), along with boxplots by year (Figure S2), season (Figure S3) and precipitations (Figure S4). To explore associations between variables, bivariate analyses were performed, including scatterplot and Pearson correlations for continuous variables and boxplots, t-test and one-way ANOVA, for categorical ones.

To reconstruct the datasets, five imputations methods were used: mean imputation, conditional mean imputation, K-Nearest Neighbor imputation, multiple imputation and Bayesian Principal Component Imputation, which are all based on multivariate regression models of $PM_{2.5}$. We built two initial regression models using log-transformed $PM_{2.5}$ and usual covariates, as previously done (Díaz-Robles et al., 2008; Koutrakis et al., 2005; Sax et al., 2007). Model 1 included meteorological and temporal covariates, as well as PM_{10} from the same monitoring station, as shown in Equation 1.

$$\begin{aligned}\ln\left(PM_{2.5}\right) = \alpha + \sum \beta pm * p_i + \sum \beta t * t_i + \sum \beta w * w_i + \sum \beta rh * rh_i + \\ + \sum \beta p * p_i + \sum \beta y * y_i + \sum \beta m * m_i + \sum \beta d * d_i + \sum \beta h * h_i + \varepsilon_i\end{aligned} \qquad \text{Equation 1}$$

Where, α is the regression intercept; *βpm*, *βt*, *βw*, *βrh*, *βp, βy*, *βm*, *βd*, and *βh* are the regression coefficients of the independent variables: $\ln(PM_{10})$, pm_i; mean temperature, t_i;

wind speed, w_i; relative humidity, rh_i; precipitations p_i; year, y_i; month, m_i; day of the week, d_i; holiday, h_i. and error term ε_i, for observation *i*. Ln (PM_{10}), mean temperature and wind speed, precipitation, and relative humidity were included as continuous variables; while year, month, day of week, and holiday were included as categorical variables, creating dummy variables for each level. Additionally, Model 2 was fitted in a similar was than Model 1, but including the logs of $PM_{2.5}$ and PM_{10} from a second monitoring site, the *Museo Ferroviario* station.

Once solved, $PM_{2.5}$ could be expressed as the product of terms representing the concentration impact factor (*f*) for each variable, which were calculated by exponentiating the estimated *βs*, as shown in Equations 2 and 3.

$$f_i = \exp^{\beta x_i} \qquad \text{Equation 2}$$

$$PM_{2.5} = \alpha \cdot f_{p,i} \cdot f_{t,i} \cdot f_{w,i} \cdot f_{rh,i} \cdot f_{p,i} \cdot f_{y,i} \cdot f_{m,i} \cdot f_{d,i} \cdot f_{h,i} \qquad \text{Equation 3}$$

With f_i being the concentration impact factor for any given regression estimate *β* for variable *x* in observation *i*; *α* being the $PM_{2.5}$ concentrations when all covariates hold their reference values; and $f_{p,i}$, $f_{t,i}$, $f_{w,i}$, $f_{rh,i}$, $f_{p,i}$, $f_{y,i}$, $f_{m,i}$, $f_{d,i}$, *and* $f_{h,i}$ being the concentration impact factors for ln(PM_{10}), temperature, relative humidity, precipitations, year, month, day of the week and holiday, respectively. Notice that a sensitivity analysis was performed using Reduced Major Axis (RMA) regression to examine the functional relationship between $PM_{2.5}$ and PM_{10}.

Subsequently, the five imputations methods were applied to reconstruct the dataset. The first method was single imputation using the mean, where missing $PM_{2.5}$ values were replaced by the mean. The second imputation method, i.e., single imputation using conditional mean, where missing $PM_{2.5}$ values were replaced by estimates from the multiple linear regression model for all observations with complete covariates data. The third method was K-Nearest Neighbor imputation. Here, we used the "mi impute pmm" command in STATA 13 (StataCorp, College Station, TX) with 20 imputation sets and the 10 nearest neighbors. The command fills in the missing data with the closest values based on the absolute difference between the linear prediction for the missing value and the complete values. The fourth method was multiple imputation and was carried out using the ‘mi’ command in STATA 13 (StataCorp, College Station, TX). Basically, multiple imputation works through two stages—the imputation stage and the analysis stage. The imputation stage creates imputations through an iterative Markov Chain Monte Carlo process, assuming a multivariate normal underlying model. Twenty imputations were executed, and each imputation iterated 2000 times, generating complete datasets for both predictors and covariates. The convergence of the algorithm was verified by examining autocorrelation and trace plots of imputed values. Each completed dataset was verified to determine if the imputation process was complete. In the analysis stage, final model parameters were estimated by combining each result using Rubin´s combination rules (StataCorp.Ltd, 2013). Finally, Bayesian Principal Component imputation was employed. The number of principal components for each model was selected. Then, an Expectation–maximization approach along with a Bayesian model was employed to calculate the likelihood for a reconstructed value (Stacklies et al., 2007) .

2.4 Validation datasets and evaluation of model performance

As we are unable to directly assess the quality of the imputation methods on missing data, a variation of a k-fold cross validation method was used (James et al., 2015). Briefly, a portion of the actual datasets was removed in a systematic way to later assess the ability of the methods to reconstruct this portion. To this end, validating datasets were built by removing $PM_{2.5}$ values from all 24 quarters from January-March, 2009 to October-December, 2014, in order to attempt to reproduce the missing pattern observed in the case study (Table S 1) . Thus, 24 sets were generated, with different quarter being removed in each set. Afterwards, each validating dataset was reconstructed using the five imputation methods and applying the two different base models (i.e., Model 1 and 2).

To evaluate the performance of environmental models, each imputed quarter was compared against the original set separately, using five indicators commonly used to assess the performance of environmental models (Bennett et al., 2013): i) Coefficient of determination (R^2), ii) Root of the mean square error (RMSE), iii) Mean Absolute Error (MAE), iv) Index of Agreement (IA), and v) Bias (B), as described in Equations 4-8:

$$R^2 = \left(\frac{\sum_{i=1}^{n} (y_i - \bar{y})(\hat{y}_i - \tilde{y})}{\sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2} \sqrt{\sum_{i=1}^{n} (\hat{y}_i - \tilde{y})^2}} \right)^2$$

Equation 4

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2}$$

Equation 5

$$MAE = \frac{1}{n} \sum_{i=1}^{n} \left| y_i - \hat{y}_i \right|$$

Equation 6

$$IA = 1 - \frac{\sum_{i-1}^{n}\left(y_i - \hat{y}_i\right)^2}{\sum_{i-1}^{n}(|\,\hat{y}_i - \bar{y}_i\,| + |\,y_i + \bar{y}_i\,|)^2} \qquad \text{Equation 7}$$

$$B = \frac{1}{n}\sum_{i-1}^{n}\ (y_i - \hat{y}_i) \qquad \text{Equation 8}$$

Where, y_i and $\hat{y}_i$ are the *i*th observation for the reconstructed and the comparison datasets, while $\bar{y}$ and $\tilde{y}$ are the means for the reconstructed and comparison datasets. R^2 is a squared version of the Pearson correlation coefficient and ranges from 0 (bad) to 1 (good). It indicates how well the model explains the variance in the observations, compared with using their mean as the prediction. RMSE expresses the error in a metric that is in the same units as the original data. MAE is similar to RMSE except that the absolute value is used instead, thus, reducing the bias towards large events. IA, in turn, resembles to the coefficient of determination but is designed to better handle differences in modeled and observed means and variances. Finally, B calculates the mean error and indicates if the model tends to under- or over-estimate the measured data with an ideal value of zero. For log-transformed variables, the exponential form informs us the relative error or bias, and can be expressed as percentage (%).

3 Results

3.1 Data completeness and pattern of missingness.

Table 1 shows data completeness for the monitoring stations. In general, completeness of PM_{10} and $PM_{2.5}$ was not very high, with losses of the order of 20%, and a slightly better performance of *Las Encinas* compared to *Museo Ferroviario*. For the other pollutants (NO_X, CO, O_3), completeness was even worse. This highlights the need to reconstruct the PM datasets, as a large portion of the health data would not have exposure data available. Meteorological variables presented a much better performance, particularly at the *Maquehue* station, so it was used for the regression models.

The pattern of missingness is shown in Table 2. When considering PM_{10}, $PM_{2.5}$, and meteorological variables (temperature, relative humidity, precipitation, and wind speed) at *Las Encinas*, the main pattern is complete case (76%), followed by missing $PM_{2.5}$ and PM_{10} (9%), and $PM_{2.5}$ only (7%) with all other patterns being negligible. A similar pattern is observed for the *Museo Ferroviario* dataset. The Little test obtained a Chi^2 of 762 (df: 72, $p<0.01$), indicating that the data seems to be MAR because there exists an identifiable pattern for the missing data. In addition, the test of missingness for independence showed that data was MAR with losses associated with other variables in the dataset: PM_{10} (OR=1.5; $p<0.01$), years (overall $p<0.01$), March (OR=0.3; $p<0.01$), April (OR=0.4; $p<0.01$), September (OR=0.5; $p=0.05$), and October (OR=0.5; $p=0.02$) (Table S4).

Table 1. Data completeness for Temuco air quality and meteorological stations.

	Pollutants										Meteorological variables											
	$PM_{2.5}$		PM_{10}		NO_X		O_3		CO		Temperature			RH			Wind speed			Precipitation		
Year	**LE**	**MF**	**LE**	**MF**	**LE**	**MF**	**LE**	**MF**	**LE**	**MF**	**LE**	**MF**	**MQ**	**LE**	**MF**	**MQ**	**LE**	**MF**	**MQ**	**LE**	**MF**	**MQ**
2009	**0.94**	**0.93**	**0.94**	**0.93**	0.69	NA	**0.94**	NA	**0.94**	NA	**0.99**	**0.99**	**1.00**	**0.94**	**0.99**	**1.00**	**0.99**	**0.91**	**1.00**	**0.99**	NA	**1.00**
2010	0.71	0.64	0.71	0.64	NA	NA	0.33	NA	0.33	NA	0.78	0.54	**1.00**	0.66	0.57	**1.00**	0.75	0.65	**1.00**	0.32	0.19	**1.00**
2011	**0.90**	0.70	**0.90**	0.70	NA	NA	0.00	NA	0.00	NA	0.89	0.72	**1.00**	**0.90**	0.71	**1.00**	0.89	0.70	**1.00**	NA	NA	**1.00**
2012	0.71	**0.98**	0.71	**0.98**	NA	NA	0.45	NA	0.45	NA	0.74	**0.98**	**1.00**	0.74	**0.98**	**1.00**	0.74	**0.94**	**1.00**	0.75	**0.98**	**1.00**
2013	0.79	0.81	0.79	0.81	NA	NA	0.44	NA	0.44	NA	0.79	0.85	**1.00**	0.75	0.73	**1.00**	0.46	0.49	**1.00**	0.45	0.50	**1.00**
2014	**0.99**	**0.98**	**0.99**	**0.98**	NA	NA	0.00	NA	0.00	NA	NA	NA	0.67	NA	NA	0.67	0.76	0.76	0.67	NA	NA	0.70
Total	0.84	0.84	0.84	0.84	0.69	NA	0.36	NA	0.36	NA	0.84	0.82	**0.95**	0.80	0.80	**0.95**	0.77	0.74	**0.95**	0.63	0.28	**0.95**

* In **bold**, completeness >90%. LE: Las Encinas. MF: Museo Ferroviario. MQ: Maquehue. NA: no available

*Wind speed: scalar average

Table 2. Missing data patterns for the *Las Encinas*, *Museo Ferroviario* and *Maquehue* monitoring stations.

Las Encinas								Museo Ferroviario							
Presence (+) / Absence (-) of data								Presence (+) / Absence (-) of data							
$PM_{2.5}$	**PM_{10}**	**Temp**	**RH**	**WS**	**PP**	**N° of days**	**% data**	**$PM_{2.5}$**	**PM_{10}**	**Temp**	**RH**	**WS**	**PP**	**N° of days**	**% data**
+	+	+	+	+	+	1675	76	+	+	+	+	+	+	1609	73
-	-	+	+	+	+	198	9	-	-	+	+	+	+	334	15
-	+	+	+	+	+	147	7	-	+	+	+	+	+	87	4
+	+	-	-	-	-	101	5	+	+	-	-	-	-	85	4
+	-	+	+	+	+	47	2	+	-	+	+	+	+	37	2
+	-	-	-	-	-	7	<1	+	-	-	-	-	-	22	1
+	+	-	-	+	+	5	<1	+	+	+	-	+	+	6	<1
+	+	+	-	+	+	5	<1	+	+	-	-	+	+	4	<1
+	+	-	-	+	-	1	<1	+	+	-	-	+	-	1	<1

Temp: temperature; RH: relative humidity; WS: wind speed; PP: precipitation

3.2 Variable characterization

Table 3 and Figure S2 show summary statistics and distributions for $PM_{2.5}$ and PM_{10}. Overall, $PM_{2.5}$ and PM_{10} concentrations exceeded national standards and international guidelines with $PM_{2.5}$ concentrations being significantly above the national annual standard of 20 µg/m^3 (Ministerio de Medio Ambiente, 2014) and the WHO annual Air Quality Guideline of 10 µg/m^3 (World Health Organization, 2006). Many days exceeded the national daily standard of 50 µg/m^3, and even reached concentrations as high as 200 µg/m^3. PM_{10} also showed concentrations above standards, but mainly driven by $PM_{2.5}$, as about 80% of PM_{10} is comprised of $PM_{2.5}$ (Ministerio del Medio Ambiente, 2015).

Table 3. Summary statistics for $PM_{2.5}$ and PM_{10}, by year and station.

	$PM_{2.5}$				PM_{10}			
	Las Encinas		Museo Ferroviario		Las Encinas		Museo Ferroviario	
Year	Mean	SD	Mean	SD	Mean	SD	Mean	SD
2009	42.4	51.6	44.0	46.0	64.3	60.6	52.5	48.1
2010	34.3	49.9	18.7	19.8	62.3	50.6	30.2	20.6
2011	47.6	44.2	49.8	46.3	65.5	54.3	74.0	55.5
2012	50.6	57.3	37.9	45.6	72.3	63.1	54.4	47.2
2013	40.4	44.0	41.5	41.5	57.3	48.2	57.5	42.1
2014	31.5	37.8	30.5	38.1	47.1	39.7	53.2	42.1
Period	40.9	47.8	37.1	42.1	61.2	53.6	54.5	46.0

SD: standard deviation.

The bivariate analyses (Figure 2) show strong associations of $PM_{2.5}$ with temporal variables such as some years (with no temporal trend) and month (higher in winter), but not with weekday or weekends. Additional associations were observed with PM_{10} (directly associated), temperature (higher when cold), relative humidity (higher when humid), and wind speed (higher when stagnant), but not with precipitations. When analyzing hourly patterns (Figure S3), highest concentrations were observed at night from 6 pm to 4 am, independent of the day of the week, with the pattern more pronounced in winter, and with little evidence of other peaks associated with traffic-rush hours. These patterns are all in agreement with small, residential wood-burning particles being the main source of $PM_{2.5}$, which persist in summer due to the use of stoves for cooking, although attenuated.

3.3 Regression model and imputation.

Results of initial regression models for log $PM_{2.5}$ of *Las Encinas* are shown in Table 4. Model 1, which included predictors from *Las Encinas* only, presented a high R^2 of 0.91, and RMSE of 0.317, implying an error of about 31%. Strong, significant predictors were PM_{10} (8% increase per each 10% of increase in PM_{10}), temperature (17% decrease per five-degree increase), and wind speed (16% decrease per 10-knots increase). Some temporal variables remained significant after controlling for pollutants and meteorology, with higher $PM_{2.5}$ in 2011 compared to other years and in winter months. Holidays and weekdays were not significant. For Model 2, which also included predictors from *Museo Ferroviario*, the R^2 increased to 0.94, and RMSE decreased to 0.262, implying a smaller error of 29%. Results were similar to Model 1 but included impacts from *Museo Ferroviario* with increases in $PM_{2.5}$ and PM_{10} being associated with increases adn decreases in $PM_{2.5}$ at *Las Encinas*, respectively. This negative coefficient for PM_{10} might be partially explained by a local source of coarse particles in *Museo Ferroviario* not present in *Las Encinas*, which can be further influenced by collinearity between variables. In general, models were in agreement with the notion that residential wood - burning is the main source of $PM_{2.5}$. Note that similar results were obtained for the sensitivity analysis of $PM_{2.5}$ and PM_{10} in both models with the RMA regression (Table S 5).

Table 4. Regression models for Ln($PM_{2.5}$) using the complete case approach.

	Model 1: Predictors from Las Encinas and Maquehue				Model 2: Predictors from Las Encinas, Museo Ferroviario and Maquehue			
	N=1657, completeness 80%, R^2=0.910, RMSE=0.317				N=1379, completeness 67%, R^2=0.941, RMSE=0.262			
Effect	**Est.**	**SE**	**p-value**	**CIF**	**Est.**	**SE**	**p-value**	**CIF**
Intercept	-0.338	0.160	**0.03**	0.71	0.005	0.150	0.97	1.01
Year			**<0.01***				**<0.01***	
2010	-0.105	0.027	**<0.01**	0.90	-0.042	0.028	0.13	0.96
2011	0.232	0.025	**<0.01**	1.26	0.188	0.025	**<0.01**	1.21
2012	-0.124	0.027	**<0.01**	0.88	-0.001	0.026	0.96	0.99
2013	-0.189	0.026	**<0.01**	0.83	-0.088	0.025	**<0.01**	0.92
2014	-0.187	0.028	**<0.01**	0.83	0.077	0.029	**0.01**	1.08
Month			**<0.01***				**<0.01***	
February	-0.100	0.039	**0.01**	0.90	-0.096	0.040	**0.02**	0.91
March	0.057	0.039	0.14	1.06	0.070	0.039	0.07	1.07
April	0.421	0.045	**<0.01**	1.52	0.306	0.045	**<0.01**	1.36
May	0.641	0.050	**<0.01**	1.90	0.420	0.049	**<0.01**	1.52
June	0.565	0.052	**<0.01**	1.76	0.326	0.053	**<0.01**	1.38
July	0.532	0.054	**<0.01**	1.70	0.334	0.053	**<0.01**	1.40
August	0.536	0.052	**<0.01**	1.71	0.361	0.050	**<0.01**	1.43
September	0.487	0.048	**<0.01**	1.63	0.413	0.046	**<0.01**	1.51
October	0.113	0.045	**0.01**	1.12	0.165	0.042	**<0.01**	1.18
November	-0.014	0.042	0.73	0.99	0.114	0.040	**<0.01**	1.12
December	-0.258	0.039	**<0.01**	0.77	-0.054	0.037	0.15	0.95
Day of the week			0.38*				0.33*	
Monday	-0.02	0.029	**0.50**	0.98	0.023	0.027	0.39	1.02
Tuesday	-0.03	0.029	0.36	0.97	0.001	0.027	0.99	1.00
Wednesday	-0.06	0.029	**0.05**	0.94	-0.023	0.027	0.40	0.98
Thursday	-0.05	0.029	0.10	0.95	-0.037	0.027	0.17	0.96
Friday	-0.05	0.029	0.09	0.95	-0.017	0.027	0.51	0.98
Saturday	-0.01	0.029	0.65	0.99	0.008	0.026	0.76	1.01
Holiday	-0.071	0.039	0.07	0.93	-0.073	0.032	**0.02**	0.93
Temperature	-0.037	0.004	**<0.01**	0.83	-0.030	0.003	**<0.01**	0.86
RH	0.009	0.001	**<0.01**	1.09	0.005	0.001	**<0.01**	1.05
Wind speed	-0.015	0.003	**<0.01**	0.86	-0.011	0.003	**<0.01**	0.90
Precipitation	-0.001	0.001	0.81	0.99	-0.002	0.001	0.11	0.99
Ln(PM_{10}), Las Encinas	0.825	0.018	**<0.01**	1.08	0.711	0.023	**<0.01**	1.07
Ln($PM_{2.5}$), Museo Ferroviario	na	na	na	na	0.499	0.023	**<0.01**	1.05
Ln(PM_{10}), Museo Ferroviario	na	na	na	na	-0.341	0.027	**<0.01**	0.97

The estimates are expressed as one-unit increase in the predictor. Reference variables are 2009, January, Sunday and working day. *Overall p-value for the variable. CIF: concentration impact factor. CIF is referred to changes in predictors of: ΔPM_{10}=10%; $\Delta PM_{2.5}$=10%; ΔTemp=5ºC; ΔWS=10knots; ΔRH=10%; na= not applicable; Wind speed: scalar average

3.4 Performance of imputation methods on validation datasets.

The results of the imputation methods on full and validation datasets are shown in Table 5, Figures S 5- S 6. In general, K-Nearest Neighbor presented a better performance than other imputations methods in both full and validation datasets. However, to the contrary of multiple imputation, K-Nearest neighbor was unable to reconstruct the full dataset because of missing values in the covariates (keeping missing data about 12%) (Figure S5). Model performance improved when including data from another station (*Museo Ferroviario,* Model 2) (Figure 3). For the full dataset, multiple imputation using model 2 provided the highest completeness (100%) with a lower error (e^{RMSE}=27%, e^{MAE}=24%), and lower bias (e^{Bias}=3.9%), thus being a promising option to reconstruct the Temuco dataset. The lower performance was observed for Bayesian principal component imputation for both models. When challenged with the validation datasets, the performance remained for most indicators and most datasets, but decreased slightly for R^2 and IA, in general, and particularly for some sets. In addition, for some sets (p25 - p75), bias was away from 0 on the order of 10%-20%, indicating that in some cases a small bias can be introduced in the set due to the imputation process.

Table 5. Results of imputation methods on validation datasets.

Model	**Obs**	**R^2**	**RMSE (%)***	**MAE(%)****	**Bias(%)*****	**IA**
Full dataset						
Model 1:						
Complete case analysis	1657	0.91	37	31	4.9	0.98
Mean Imputation	1804	0.85	49	33	2.3	0.96
Conditional Mean Imputation	1804	0.92	36	31	4.9	0.98
K-Nearest Neighbor	1804	0.91	25	25	2.1	0.98
Multiple Imputation	2061	0.91	34	31	5.8	0.99
Bayesian Principal component analysis	2061	0.86	45	37	8.1	0.96
Model 2:						
Complete case analysis	1379	0.94	30	24	3.2	0.98
Mean Imputation	1439	0.91	38	25	1.2	0.98
Conditional Mean Imputation	1439	0.94	29	24	3.2	0.99
K-Nearest Neighbor	1439	0.94	25	25	2.1	0.98
Multiple Imputation	2061	0.94	27	24	3.9	0.98
Bayesian Principal component analysis	2061	0.89	40	32	6.1	0.97
Validation datasets	Median	Median	Median	Median	Median	Median
	(p25-p75)	(p25-p75)	(p25-p75)	(p25-p75)	(p25-p75)	(p25-p75)
Model 1:						
Mean Imputation	80	0.80	26	28	2.9	0.92
	(63-88)	(0.46-0.90)	(19-28)	(24-34)	(-7.4-16.4)	(0.76-0.96)
Conditional Mean Imputation	80	0.80	27	28	4.3	0.91
	(63-88)	(0.45-0.89)	(21-30)	(22-32)	(-12.5-9.8)	(0.78-0.97)
K-Nearest Neighbor	80	0.80	28	28	4.1	0.90
	(63-88)	(0.45-0.89)	(21-30)	(22-32)	(-12.1-9.6)	(0.78-0.97)
Multiple Imputation	82.5	0.78	29	33	7.7	0.87
	(66-10)	(0.41-0.89)	(21-33)	(27-43)	(-21.9-17.4)	(0.72-0.95)
Bayesian Principal component analysis	82	0.75	29	40	9.2	0.89
	(65-89)	(0.37-0.84)	(21-31)	(30-54)	(-19.9-32.0)	(0.62-0.92)
Model 2:						
Mean Imputation	71.5	0.83	20	22	0.6	0.95
	(25-82)	(0.73-0.91)	(18-24)	(18-26)	(-7.8-6.5)	(0.92-0.97)
Conditional Mean Imputation	72	0.85	21	22	-1.8	0.95
	(25-82)	(0.74-0.91)	(19-26)	(19-27)	(-7.8-5.6)	(0.91-0.97)
K-Nearest Neighbor	80	0.80	28	28	4.1	0.90
	(63-88)	(0.45-0.89)	(21-30)	(22-32)	(-12.1-9.6)	(0.78-0.97)
Multiple Imputation	83	0.81	26	31	-2.8	0.92
	(66-90)	(0.61-0.90)	(20-31)	(22-37)	(-2.8--13.8)	(0.81-0.96)
Bayesian Principal component analysis	82	0.79	25	37	5.2	0.89
	(65-89)	(0.55-0.86)	(20-31)	(24-51)	(-19.9-24.7)	(0.69-0.94)

Obs: Observations; RMSE: Root mean square error; MAE: Mean absolute error, IA: Index of agreement .
*RMSE(%)=[exp(RMSE)-1]*100; **MAE(%)=[exp(MAE)-1]*100; ***Bias(%)=[exp(Bias)-1]*100

4 Discussion

In this article, we attempted to reconstruct the $PM_{2.5}$ dataset from Temuco, a mid-size city heavily impacted by residential wood-burning. As with in many cities in Chile, the dataset presented a high rate of losses (over 20%), which could jeopardize further health analysis. Data seemed to be MAR with some associations with other variables, but in agreement with losses due to technical failures. Regression models were successful in predicting $PM_{2.5}$ with many predictors, such as temperature and season associated with residential wood-burning (Jorquera et al., 2018), and with better performance when including data from another station (*Museo Ferroviario*).

When applying imputation methods, multiple imputation was able to reconstruct the dataset with improved performance when including covariates from the other station. The performance seemed promising in terms of R^2, errors and bias, even when challenged with validation datasets. K-Nearest Neighbor showed slightly better performance than multiple imputation for error and bias but was not able to reconstruct the full dataset. The lower performance of multiple imputation is expected as it incorporates the imputation error (Rubin, 1996).

Rather few previous studies have used imputation methods to reconstruct datasets. In a comprehensive study using data with missingness near 25% from Helsinski, Finland, and Belfast, North Ireland; similar measures of performance were found with R^2 of 0.49, RMSE of 0.22 and MAE of 0.16 (Junninen et al., 2004). Additionally, they found that single imputation methods underestimated the error variance and accuracy of missing data compared to multiple imputation, which might explain our results. In another study using datasets in La Coruña, Spain, several imputation methods were compared (Gómez-Carracedo et al., 2014). They used factor analysis with Varimax rotation along with the imputation methods, but did not provide overall performance measures, in terms of completeness, error, and bias, and did not challenge the methods with validation sets. They found that multiple imputation had more scattered results when datasets had more than 43.5% of missingness and were poorly correlated with other variables; however, results were similar when missingness was medium, as in our case. Finally, an infant cohort study investigating the effects of pollution on asthma risk (Roda et al., 2014), compared methods for imputing indoor domestic pollutants. The complete case reduced the statistical power, while single imputation overestimated the association and multiple imputation was too conservative and unable to show significant associations. Considering this experience, it seems necessary that researchers continue attempting the reconstruction of datasets, particularly where more needed, such as low- middle-income countries and small cities. It seems important to provide overall indicators of performance, as these can be locally driven by the quality of the data and the base regression model. Junger and de Leon (2015) developed a time-series for an air pollution simulation study using complete case analysis, unconditional mean imputation, conditional mean imputation and other approaches such as a regular Expectation Maximization algorithm (EM), EM algorithm filtered by splines, among others. They found that when the amount of missing data was less than 5%, the complete case analysis had a good performance. However, when the missing data was higher the validity of estimates degraded.

The results are limited only to Temuco and for the time-period under study. The combination of explanatory variables selected in our imputation models for Temuco might differ in other locations. For instance, the application of this framework to areas located near large industrial complexes or surface mining operation might highlight wind direction to be a strong predictor for ambient $PM_{2.5}$, whereas the model for Temuco did not include this variable in the final model. Similarly, cities located in arid regions have a larger influence from coarse particles, weakening the correlation between PM_{10} and $PM_{2.5}$. However, the methodological framework employed in this study to identify the best imputation model could be usefully replicated in other regions and cities. Therefore, it would be interesting to extend the current approach to other time periods in Temuco, other cities in Chile and elsewhere, taking into in consideration the specific atmospheric composition, sources and dynamics of the air shed in individual cities.

A limitation of this work is the fact that the background concentration of air pollution or the boundary layer are not measured by the monitoring air quality network and could not be included in the statistical models. However, previous research in the study area have shown that the main source or air pollution is residential wood burning (Jorquera et al., 2018; SICAM, 2014; Villalobos et al., 2017, 2015). A potential limitation of using imputation methods to predict missing values would occur in the case that the data were MNAR, as it might introduce bias in the data set. Results from our validation dataset, showed small bias in general, but more significant in some specific cases like Bayesian principal component analysis. This is a warning as in some circumstances a bias in $PM_{2.5}$ estimation might be introduced even if the MAR assumptions would be met; however, this bias seems not to be high, on the order of 10%-20%. In any circumstance, the possibility of biasing the health estimates due to the introduction of a small bias during the imputation process should be weighed against the possible bias incurred by not including the full dataset in the analysis.

In summary, our results show that using imputation methods, particularly multiple imputation, can be to a certain extent successful in reconstructing an air quality data set with relatively low-medium missingness in a real-life situation. This is relevant for datasets in small locations where the problem of missing data might be more frequent alongside with serious environmental health problems.

Acknowledgement

This work was supported as part of the project: “Impact of Wood Burning Air Pollution on Preeclampsia and other Pregnancy Outcomes in Temuco, Chile” (DPI20140093) by CONICYT and Research Councils UK. Juana Maria Delgado-Saborit is supported by the European Union’s Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 750531. María Elisa Quinteros was supported by a doctoral scholarship by CONICYT Beca Doctorado Nacional No 21150801, Chile. We acknowledge Xavier Basagaña for his technical help, Payam Dadvand for his intellectual assistance, Gloria Icaza Noguera for reviewing the manuscript, and Estela Blanco for her help in reviewing English writing of the article.

Competing financial interests

The authors declare they have no competing interests.

5 References

Bennett, N.D., Croke, B.F.W.W., Guariso, G., Guillaume, J.H.A.A., Hamilton, S.H., Jakeman, A.J., Marsili-Libelli, S., Newham, L.T.H.H., Norton, J.P., Perrin, C., Pierce, S.A., Robson, B., Seppelt, R., Voinov, A.A., Fath, B.D., Andreassian, V., 2013. Characterising performance of environmental models. Environ. Model. Softw. 40, 1–20. https://doi.org/10.1016/j.envsoft.2012.09.011

Bishop, C.M., 1999. Variational principal components. IEE Conf. Publ. Artif. Neural Networks 509–514.

Díaz-Robles, L.A., Ortega, J.C., Fu, J.S., Reed, G.D., Chow, J.C., Watson, J.G., Moncada-Herrera, J.A., 2008. A hybrid ARIMA and artificial neural networks model to forecast particulate matter in urban areas: The case of Temuco, Chile. Atmos. Environ. 42, 8331–8340. https://doi.org/10.1016/j.atmosenv.2008.07.020

Dirección Metereológica de Chile, 2016. Climatologia. Available from http://www.meteochile.cl/PortalDMC-web/index.xhtml.

Dixon, J.K., 1979. Pattern recognition with partly missing data. IEEE Trans. Syst. Man, Cybern. 10 617–621.

Gómez-Carracedo, M.P., Andrade, J.M., López-Mahía, P., Muniategui, S., Prada, D., 2014. A practical comparison of single and multiple imputation methods to handle complex missing data in air quality datasets. Chemom. Intell. Lab. Syst. 134, 23–33. https://doi.org/10.1016/j.chemolab.2014.02.007

Gómez, W., Salgado, H., Vásquez, F., Chávez, C., 2014. Using stated preference methods to design cost-effective subsidy programs to induce technology adoption: An application to a stove program in southern Chile. J. Environ. Manage. 132, 346–357. https://doi.org/10.1016/j.jenvman.2013.11.020

Green, J., Sánchez, S., 2012. La Calidad del Aire en América Latina: Una Visión Panorámica. Clean Air Institute. Available from http//www.minambiente.gov.co/images/AsuntosambientalesySectorialyUrbana/pdf/contaminacion_atmosferica/La_Calidad_del_Aire_en_Am%C3%A9rica_Latina.pdf.

Greenland, S., Finkle, W.D., 1995. A critical look at methods for handling missing covariates in epidemiologic regression analyses. Am. J. Epidemiol. 142, 1255–1264. https://doi.org/https://doi.org/10.1093/oxfordjournals.aje.a117592

INE, 2017. Anuarios parque de vehículos en circulación. Available from http//historico.ine.cl/canales/chile_estadistico/estadisticas_economicas/transporte_y_comunicaciones/parquevehiculos.php.

James, G., Witten, D., Hastie, T., Tibshirani, R., 2015. Resampling methods, in: An Introduction to Statistical Learning. pp. 176–184.

Jorquera, H., Barraza, F., Heyer, J., Valdivia, G., Schiappacasse, L.N., Montoya, L.D., 2018. Indoor PM2.5in an urban zone with heavy wood smoke pollution: The case of Temuco, Chile. Environ. Pollut. 236, 477–487. https://doi.org/10.1016/j.envpol.2018.01.085

Junger, W., de Leon, A.P., 2009. Missing Data Imputation in Time Series of Air Pollution. Epidemiology 20. https://doi.org/10.1097/01.ede.0000362970.08869.60

Junger, W.L., de Leon, A.P., 2015. Imputation of missing data in time series for air pollutants.

Atmos. Environ. 102, 96–104. https://doi.org/https://doi.org/10.1016/j.atmosenv.2014.11.049

Junninen, H., Niska, H., Tuppurainen, K., Ruuskanen, J., Kolehmainen, M., 2004. Methods for imputation of missing values in air quality data sets. Atmos. Environ. 38, 2895–2907. https://doi.org/https://doi.org/10.1016/j.atmosenv.2004.02.026

Klebanoff, M.A., Cole, S.R., 2008. Use of multiple imputation in the epidemiologic literature. Am. J. Epidemiol. 168, 355–357. https://doi.org/10.1093/aje/kwn071

Koutrakis, P., Sax, S.N., Sarnat, J. a, Coull, B., Demokritou, P., Oyola, P., Garcia, J., Gramsch, E., 2005. Analysis of PM10, PM2.5, and PM2 5-10 concentrations in Santiago, Chile, from 1989 to 2001. J. Air Waste Manag. Assoc. 55, 342–351. https://doi.org/10.1080/10473289.2005.10464627

Little, R., Rubin, D., 1987. Statistical Analysis With Missing Data, 2nd ed. Wiley Interscience, Hoboken, NJ.

Little, R.J.A., 1988. A Test of Missing Completely at Random for Multivariate Data with Missing Values. J. Am. Stat. Assoc. 83, 1198–1202. https://doi.org/10.2307/2290157

Ministerio de Desarrollo Social, 2011. Encuesta Caracterización Socio económica. Perfil Región de la Araucanía. Available from http//observatorio.ministeriodesarrollosocial.gob.cl/casen/casen_perfil_9.php.

Ministerio de Medio Ambiente, 2017. Sistema de Información Nacional de Calidad del Aire. Región La Araucanía Estac. Monit. la Calid. del aire. Available from http//sinca.mma.gob.cl/index.php/region/index/id/IX.

Ministerio de Medio Ambiente, 2014. Planes de Descontaminación Atmosférica Estrategia 2014 - 2018. Available from http//portal.mma.gob.cl/planes-de-descontaminacion-atmosferica-estrategia-2014-2018/.

Ministerio del Medio Ambiente, 2018. Normativa aplicable - Sistema de Información Nacional de Calidad del Aire. Gob. Chile, https://si.

Ministerio del Medio Ambiente, 2015. Plan de prevención y descontaminación atmosférica Temuco y Padre Las Casas. DS 8 del 2015 MMA.

Minsal, 2016. Diagnosticos regionales en salud con enfoque en determinantes sociales. Ficha regional: Araucania. Available from http//epi.minsal.cl/datos-drs/9_araucania.pdf.

Molina Sepúlveda, V., Oyarzo Gómez, E., 2013. Estudio de la factibilidad de un sistema eficiente de calefacción para la ciudad de Temuco. Available from http//cybertesis.uach.cl/tesis/uach/2013/bpmfem722e/doc/bpmfem722e.pdf.

Pascal, M., Corso, M., Chanel, O., Declercq, C., Badaloni, C., Cesaroni, G., Henschel, S., Meister, K., Haluza, D., Martin-Olmedo, P., Medina, S., Aphekom group, 2013. Assessing the public health impacts of urban air pollution in 25 European cities: Results of the Aphekom project. Sci. Total Environ. 449, 390–400. https://doi.org/10.1016/j.scitotenv.2013.01.077

Riojas-Rodriguez, H., da Silva, A.S., Texcalac-Sangrador, J.L.J.L., Moreno-Banda, G.L., Riojas-Rodríguez, H., Silva, A.S. da, Texcalac-Sangrador, J.L.J.L., Moreno-Banda, G.L., 2016. Air pollution management and control in Latin America and the Caribbean: implications for climate change. Rev. Panam. Salud Publica 40, 150–159.

Roda, C., Nicolis, I., Momas, I., Guihenneuc, C., 2014. New insights into handling missing values in environmental epidemiological studies. PLoS One 9.

https://doi.org/10.1371/journal.pone.0104254

Rubin, D., 1987. Multiple Imputation for Nonresponse in Surveys. John Wiley & Sons, New York. https://doi.org/10.1002/9780470316696

Rubin, D.B., 1996. Multiple Imputation after 18+ Years. J. Am. Stat. Assoc. https://doi.org/10.1080/01621459.1996.10476908

Ruiz-Rudolph, P., 2014. Impact of Wood Burning Air Pollution on Preeclampsia and other Pregnancy Outcomes in Temuco, Chile (DPI20140093). CONICYT and Research Councils UK.

Sax, S.N., Koutrakis, P., Ruiz Rudolph, P.A., Cereceda-Balic, F., Gramsch, E., Oyola, P., 2007. Trends in the elemental composition of fine particulate matter in Santiago, Chile, from 1998 to 2003. J. Air Waste Manag. Assoc. 57, 845–855. https://doi.org/10.3155/1047-3289.57.7.845

Schafer, J.L., Graham, J.W., 2002. Missing data: Our view of the state of the art. Psychol. Methods 7, 147–177. https://doi.org/10.1037//1082-989X.7.2.147

SICAM, 2014. Emission Inventory for the Temuco-Padre Las Casas Metropolitan Area: Year 2013: Residential Wood Burning. Temuco.

Stacklies, W., Redestig, H., Scholz, M., Walther, D., Selbig, J., 2007. pcaMethods – a Bioconductor package providing PCA methods for incomplete data. Bioinformatics 23, 1164–1167.

StataCorp.Ltd, 2013. Stata Multiple-Imputation Reference Manual, Publication, A Stata Press. https://doi.org/10.1016/j.enpol.2012.08.024

Sterne, J.A.C., White, I.R., Carlin, J.B., Spratt, M., Royston, P., Kenward, M.G., Wood, A.M., Carpenter, J.R., 2009. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ 338, b2393. https://doi.org/10.1136/bmj.b2393

Stuart, E.A., Azur, M., Frangakis, C., Leaf, P., 2009. Multiple imputation with large data sets: A case study of the children’s mental health initiative. Am. J. Epidemiol. 169, 1133–1139. https://doi.org/10.1093/aje/kwp026

Toro A., R., Campos, C., Molina, C., Morales S., R.G.E., Leiva-Guzmán, M.A., Toro A, R., Campos, C., Molina, C., Morales S, R.G.E., Leiva-Guzman, M.A., Toro A., R., Campos, C., Molina, C., Morales S., R.G.E., Leiva-Guzmán, M.A., 2015. Accuracy and reliability of Chile’s National Air Quality Information System for measuring particulate matter: Beta attenuation monitoring issue. Environ. Int. 82, 101–109. https://doi.org/10.1016/j.envint.2015.02.009

van Buuren, S., 2012. Flexible Imputation of Missing Data. CRC Press (Chapman & Hall).

Villalobos, A.M., Barraza, F., Jorquera, H., Schauer, J.J., 2017. Wood burning pollution in southern Chile: PM2.5 source apportionment using CMB and molecular markers. Environ. Pollut. 225, 514–523. https://doi.org/10.1016/j.envpol.2017.02.069

Villalobos, A.M., Barraza, F., Jorquera, H., Schauer, J.J., 2015. Chemical speciation and source apportionment of fine particulate matter in Santiago, Chile, 2013. Sci. Total Environ. 512–513, 133–142. https://doi.org/10.1016/j.scitotenv.2015.01.006

World Health Organization, 2016. WHO Global Urban Ambient Air Pollution Database (update2016). Avalaible from http//www.who.int/phe/health_topics/outdoorair/databases/cities/en/.

World Health Organization, 2006. Air Quality Guidelines. Global update 2005. Available from http//www.euro.who.int/__data/assets/pdf_file/0005/78638/E90038.pdf.

Figure 1

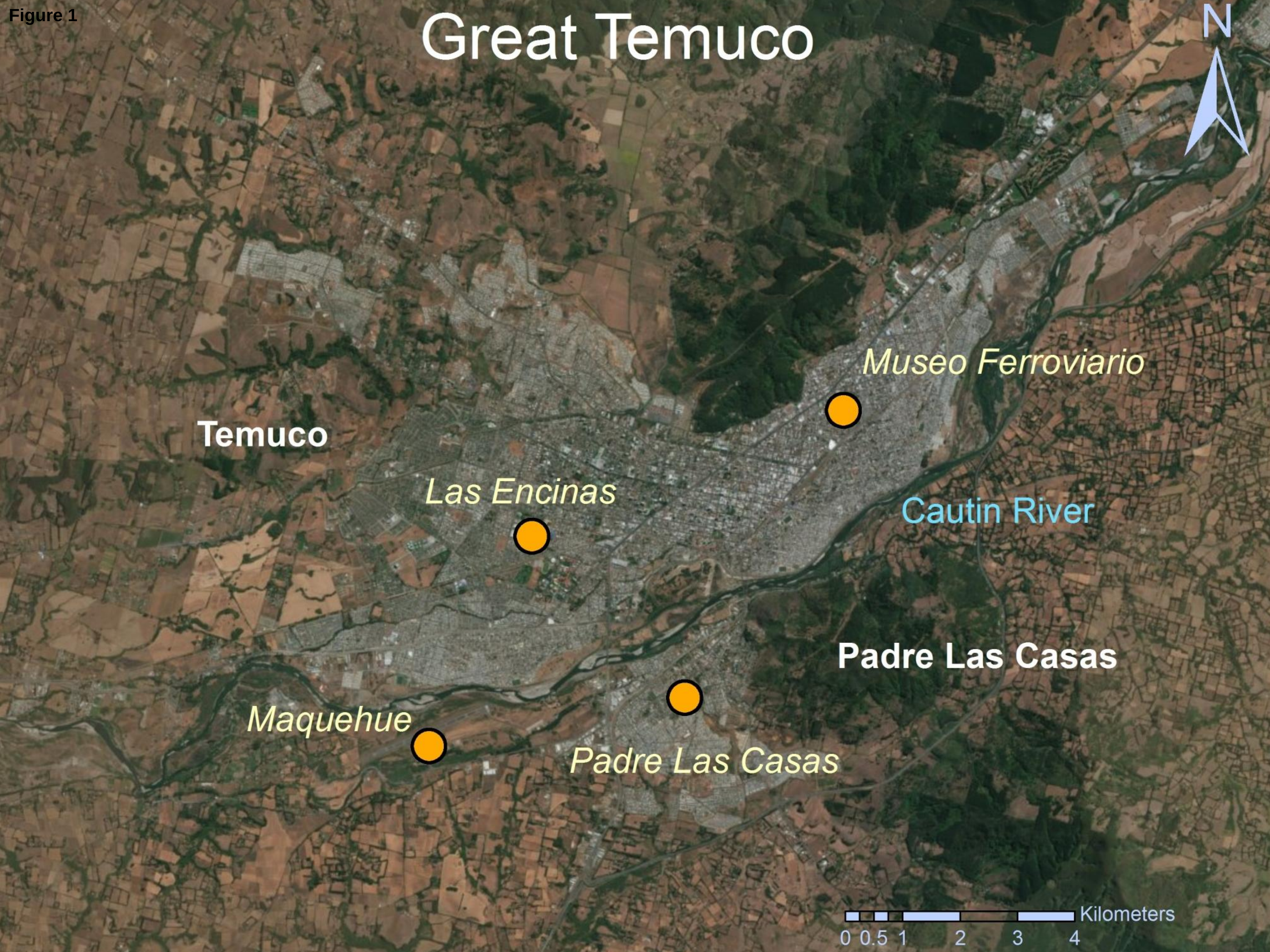
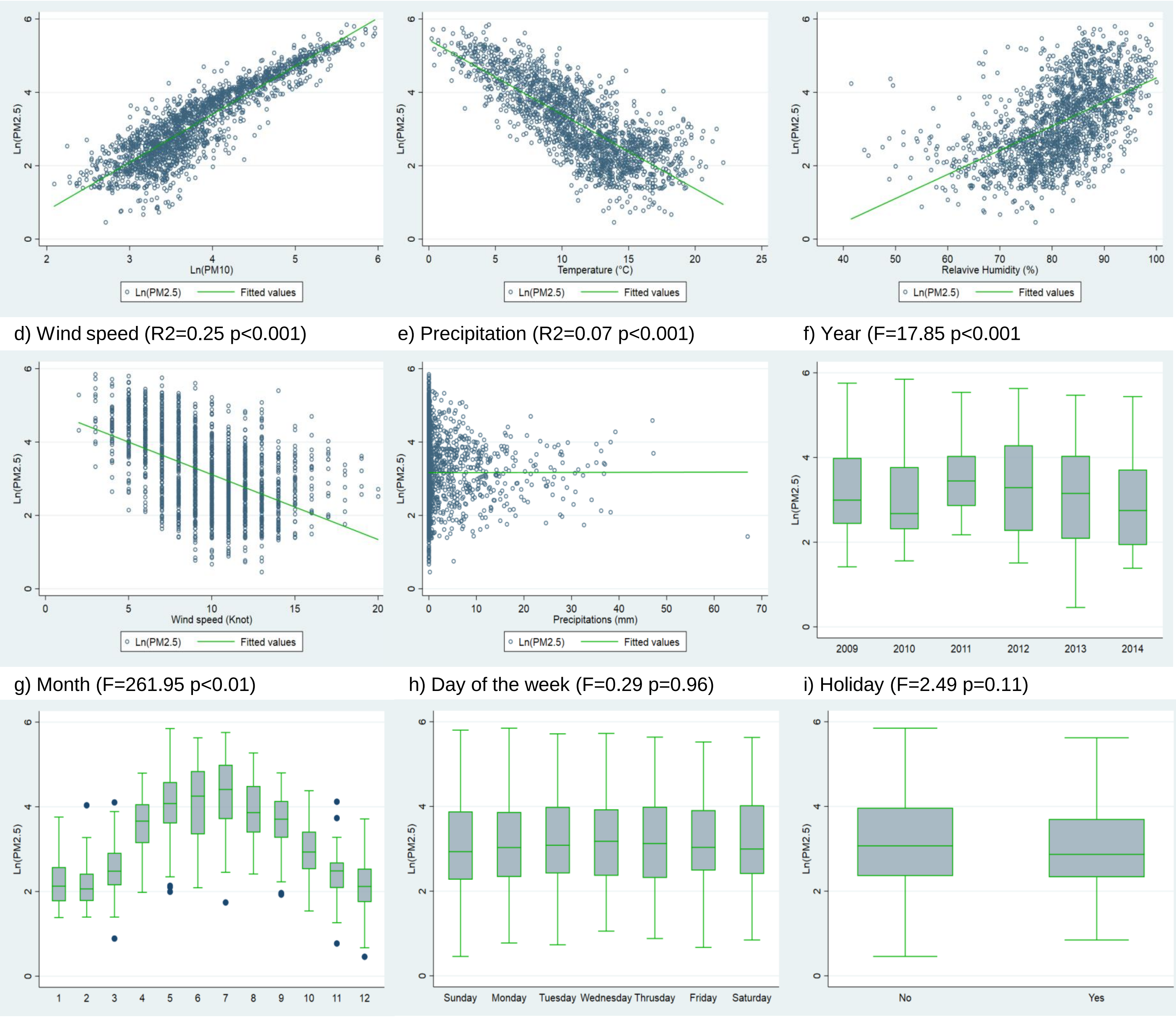

Great Temuco
N
Museo Ferroviario
Temuco
Las Encinas
Cautin River
Padre Las Casas
Maquehue
Padre Las Casas
Kilometers
0 0.5 1 2 3 4

Figure 2



a) PM10 (R2=0.79, p<0.001)

b) Temperature (R2=0.60 p<0.001)

c) Relative humidity (R2=0.30 p<0.001)

d) Wind speed (R2=0.25 p<0.001)

e) Precipitation (R2=0.07 p<0.001)

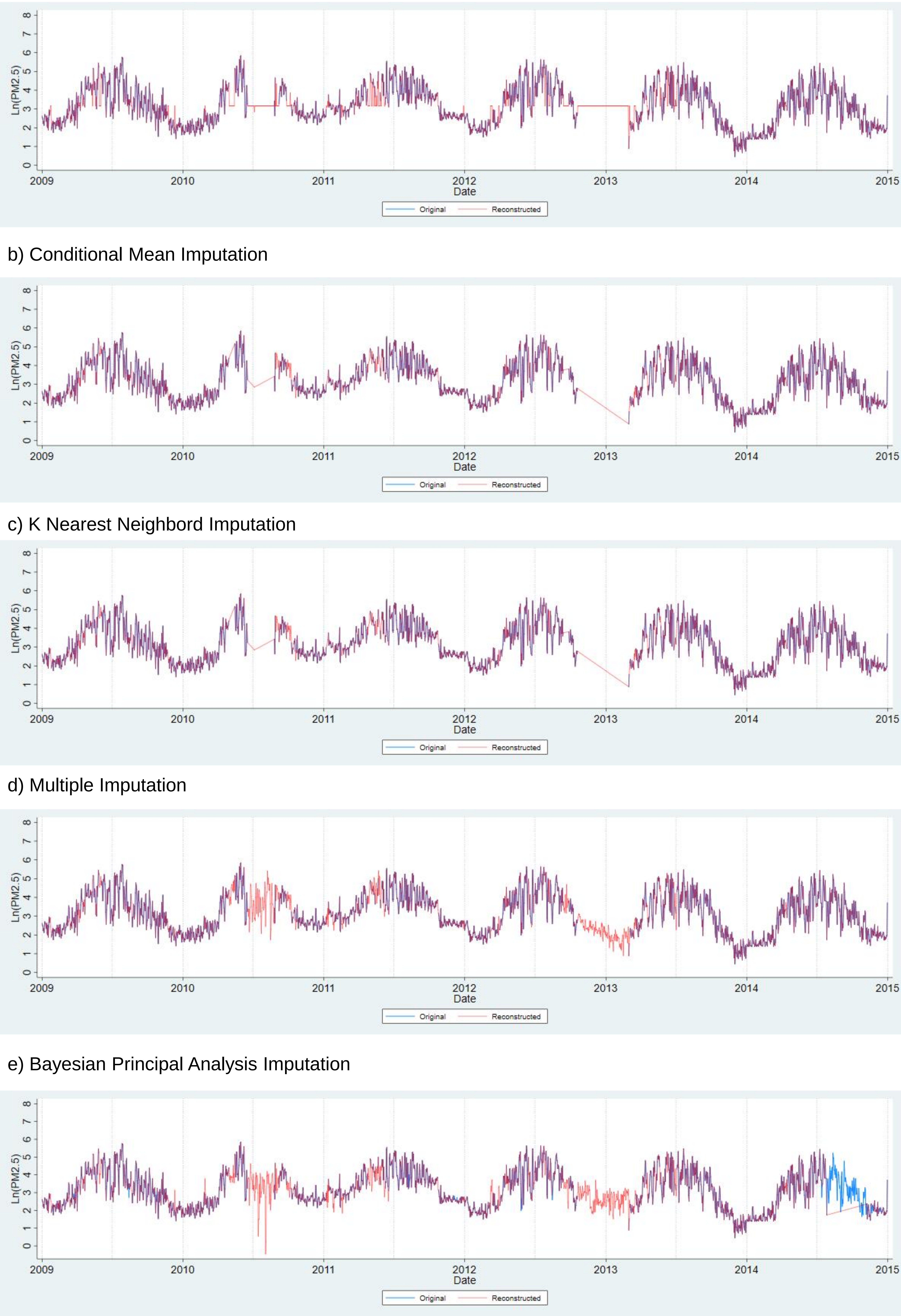
f) Year (F=17.85 p<0.001

g) Month (F=261.95 p<0.01)

h) Day of the week (F=0.29 p=0.96)

i) Holiday (F=2.49 p=0.11)

Figure 3



a) Mean Imputation

b) Conditional Mean Imputation

c) K Nearest Neighbord Imputation

d) Multiple Imputation

e) Bayesian Principal Analysis Imputation