

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](https://doi.org/10.1016/j.atmosenv.2018.11.053)

License:

Creative Commons: Attribution-NonCommercial-NoDerivs (CC BY-NC-ND)

Document Version

Peer reviewed version

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>

[Link to publication on Research at Birmingham portal](#)

General rights

Unless a licence is specified above, all rights (including copyright and moral rights) in this document are retained by the authors and/or the copyright holders. The express permission of the copyright holder must be obtained for any use of this material other than for purposes permitted by law.

- Users may freely distribute the URL that is used to identify this publication.
- Users may download and/or print one copy of the publication from the University of Birmingham research portal for the purpose of private study or non-commercial research.
- User may use extracts from the document in line with the concept of 'fair dealing' under the Copyright, Designs and Patents Act 1988 (?)
- Users may not further distribute the material nor use it for the purposes of commercial gain.

Where a licence is displayed above, please note the terms and conditions of the licence govern your use of this document.

When citing, please reference the published version.

Take down policy

While the University of Birmingham exercises care and attention in making items available there are rare occasions when an item has been uploaded in error or has been deemed to be commercially or otherwise sensitive.

If you believe that this is the case for this document, please contact UBIRA@lists.bham.ac.uk providing details and we will remove access to the work immediately and investigate.

1 A paper to be submitted to *Atmospheric Environment*

2

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

5

6 María Elisa Quinteros^{a, b}, Siyao Lu^c, Carola Blazquez^d, Juan Pablo Cárdenas-R^e,

7 Ximena Ossa^f, Juana-María Delgado-Saborit^{g,h,i, j}, Roy M. Harrison^{g,k}, Pablo Ruiz-

8 Rudolph^l

9

10 ^a Programa Doctorado en Salud Pública, Instituto de Salud Poblacional, Facultad de
11 Medicina, Universidad de Chile, Independencia 939, Independencia, Santiago, Chile.

12 ^b Departamento de Salud Pública. Facultad de Ciencias de la Salud, Universidad de
13 Talca, Avenida Lircay s/n, Talca, Chile.

14 ^c Department of Environmental Health Sciences, University of Michigan, 1415
15 Washington Heights, Ann Arbor, MI 48109, EE. UU.

16 ^d Department of Engineering Sciences, Universidad Andres Bello, Quillota 980, Viña del
17 Mar, 2531015, Chile.

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

21 ^f Departamento de Salud Pública y Centro de Excelencia CIGES, Universidad de la
22 Frontera, Caro Solar 115, Temuco, Chile.

23 ^g Division of Environmental Health and Risk Management, School of Geography, Earth
24 and Environmental Sciences, University of Birmingham, Edgbaston Birmingham
25 B152TT, UK.

26 ^h ISGlobal Barcelona Institute for Global Health, Barcelona Biomedical Research Park,
27 Doctor Aiguader 88, 08003, Barcelona, Spain.

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

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

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

33 ^l Programa de Salud Ambiental, Instituto de Salud Poblacional, Facultad de Medicina,
34 Universidad de Chile, Independencia 939, Independencia, Santiago, Chile.

35 *Corresponding author

36

37 **Corresponding Author**

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

41

42

43

44

45

46

47 **List of tables**

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

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

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

52 Table 4. Regression models for $\ln(PM_{2.5})$ using complete case approach.

53 Table 5. Results of imputation methods on validation datasets.

54

55 **List of figures**

56 Figure 1. Map of Temuco and monitoring stations.

57 Figure 2. Graphical associations between $PM_{2.5}$ from *Las Encinas* and covariates.

58 Figure 3. Reconstructed $\ln(PM_{2.5})$ concentrations in *Las Encinas* using imputation
59 methods with model 2.

60

61 **List of supplemental tables**

62 Table S 1. Bocks of missing patterns.

63 Table S 2. Summary statistics for $PM_{2.5}$ and PM_{10} at Las Encinas, by day of the week.

64 Table S 3. Summary statistics for $PM_{2.5}$ and PM_{10} at Las Encinas, by season.

65 Table S 4. Single logistics regressions of missing values against each predictor.

66 Table S 5. Sensitivity analysis results.

67

68

69

70

71 **List of supplemental figures**

72 Figure S 1. Distribution of $PM_{2.5}$ and PM_{10} at *Las Encinas* monitoring station.

73 Figure S 2. $PM_{2.5}$ and PM_{10} distributions by air quality station and year.

74 Figure S 3. Hourly $PM_{2.5}$ distribution by season.

75 Figure S 4. Precipitations distribution by year.

76 Figure S 5. Scatter plot of observed and predict $\ln(PM_{2.5})$ concentrations at *Las*

77 *Encinas* station using imputation methods with model 1.

78 Figure S 6. Scatter plot of observed and predict $\ln(PM_{2.5})$ concentrations at *Las*

79 *Encinas* station using imputation methods with model 2.

80 Figure S 7. Reconstructed $\ln(PM_{2.5})$ concentrations at *Las Encinas* monitoring station

81 using imputation methods with model 1.

82

83

84

85

86

87

88

89

90

91

92

93

94

95 **Abstract**

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

115

116 **Keywords:**

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

119 **1 Introduction**

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

130

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

141

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

151
152 Since the mid-eighties more sophisticated approaches have been introduced, including
153 expectation maximization, weighted estimating equation methods, and particularly, K-
154 Nearest Neighbor, multiple imputation and imputation using Bayesian principal
155 component analysis. The nearest neighbor imputation draws imputed values from the
156 closeness observation based on the absolute difference between the linear prediction for
157 the missing value and that for the complete values (Dixon, 1979). Multiple imputation is
158 based on Bayesian methods, and its main purpose is to properly reproduce the
159 variance/covariance matrix had the data been complete, thus providing valid inference
160 under MAR assumptions (Little and Rubin, 1987). It uses an iterative form of stochastic
161 imputation, creating multiples copies of the database, where missing values are
162 replaced by imputed values from a posterior predictive distribution using the partially
163 observed data. Subsequently, every database is analyzed and results are combined,
164 including standards errors. Therefore, data uncertainty is incorporated in the process
165 (Little and Rubin, 1987; Rubin, 1987). The Bayesian principal component analysis

166 imputation involves Bayesian estimation of missing values with the iterative expectation
167 maximization algorithm. This analysis is based on three processes: principal component
168 regression, Bayesian estimation, and an expectation–maximization (EM)-like repetitive
169 algorithm (Bishop, 1999).

170
171 Despite the fact that imputations tools are available in many statistical packages, they
172 are not often used very in epidemiological studies (Klebanoff and Cole, 2008; Sterne et
173 al., 2009; Stuart et al., 2009). Moreover, in environmental epidemiology the most
174 common approaches have been to ignore them (i.e., the complete case analysis), to
175 replace missing data based on prior knowledge, or to use single imputation, for instance,
176 from a multiple regression (Roda et al., 2014). Some studies have included multiple
177 imputation applied to air quality datasets (Junger and de Leon, 2009, 2015; Junninen et
178 al., 2004; Roda et al., 2014), but overall its application remains scarce with few tests of
179 performance in real-life situations and providing little guidance with respect to the
180 application in other settings.

181
182 Here, we propose to use imputation methods as a tool to reconstruct air quality datasets
183 and applying them to an air quality dataset in Temuco, a mid-size city in Chile as a case-
184 study. Temuco resembles the problems faced in many small-medium cities in the world,
185 whose datasets may be fragmented. It also faces a major environmental health problem
186 being heavily impacted by residential wood-burning, as many southern Chilean cities,
187 highlighting the importance of having full data for epidemiological studies (Díaz-Robles
188 et al., 2008; Gómez et al., 2014; Villalobos et al., 2017). In this study, we attempt to
189 reconstruct the database comparing five approaches: mean imputation, conditional

190 mean imputation, K-Nearest Neighbor imputation, multiple imputation and Bayesian
191 Principal Component Analysis imputation. The overall approach considers i) developing
192 a standard regression model of $PM_{2.5}$ using available predictors that could explain the air
193 pollutant concentration in the case study (i.e. meteorological and co-pollutants), ii) based
194 on the best models, applying the imputation methods to complete the datasets, iii)
195 building validation datasets by artificially removing data, and iv) assessing the
196 performance of the methods in reconstructing the removed data in the validation sets.
197 The application of the best method is expected to be used in a real-life situation in
198 Temuco by completing the $PM_{2.5}$ datasets required to build a land-use regression model,
199 which will later be used to estimate exposures in a health study of wood-burning air
200 pollution and pregnancy outcomes (Ruiz-Rudolph, 2014).

201 **2 Methods**

202 *2.1 Study Area*

203 Temuco is a mid-size city of 290,000 inhabitants located in the Araucanía Region, in
204 southern Chile (longitude 39.7°E; latitude 73.0°S) in a valley crossed by the Cautín river
205 and surrounded by hills, native forest, and agricultural fields (Minsal, 2016). The “Great
206 Temuco” is a conurbation of two cities: Temuco, to the north, and Padre Las Casas, to
207 the south across the river (Figure 1). Temuco, and the Araucanía region in general,
208 present a population of medium to low socioeconomic status, which is reflected by the
209 22.9% of the households that are classified as poor, and by the only 8.2 years of
210 schooling on average of the head of the household (Ministerio de Desarrollo Social,
211 2011). The city experiences a Mediterranean climate with oceanic influence (Csb), with
212 average temperatures close to 12°C, rainfall above 1,000 mm per year, and marked
213 seasonal differences, with cold, humid winters, and low wind speeds associated with
214 poor air pollution dispersion (Ministerio de Medio Ambiente, 2014).

215
216 The study area has some characteristics different from other many Chilean cities but
217 similar to many in the south. For example, the industrial activity in the area is low with
218 agriculture being the main economic activity (Minsal, 2016). Known air pollution sources
219 include some stationary emissions such as industrial wood- and coal-fired boilers
220 associated with the processed woods industry (Ministerio del Medio Ambiente, 2015),
221 and a medium-sized fleet of 67,800 motorized vehicles (INE, 2017). However, the
222 largest aggregated source of PM_{2.5} and PM₁₀ is the residential wood-burning that is used
223 throughout the city in winter for heating and cooking. More than 88% of homes have

224 wood-stoves, and approximately 654,000 m³ of wood are used per year (Gómez et al.,
225 2014; Ministerio del Medio Ambiente, 2015; Molina Sepúlveda and Oyarzo Gómez,
226 2013; Villalobos et al., 2017).

227 2.2 Data sources

228 The Great Temuco has an air pollution monitoring network that measures PM₁₀, PM_{2.5},
229 SO₂, NO_x, O₃, CO, and meteorological variables. This network is run by the Ministry of
230 the Environmental, and hourly data is available online (Ministerio de Medio Ambiente,
231 2017). The network is comprised by two stations in Temuco (*Las Encinas* and *Museo*
232 *Ferroviario* stations) and another one in Padre Las Casas comprise the network (Figure
233 1). The three stations began PM_{2.5} measurements in 2009. Since *Las Encinas* contains
234 more the complete sets, we focus in reconstructing its full series of PM_{2.5} from 2009 to
235 2014, so it can be later used to estimate historical exposures. Note that there is no
236 available dataset capturing the regional contribution of air pollutants levels in the studied
237 area. Additional meteorological data were obtained from the *Maquehue* station run by
238 the Meteorological Office of Chile (Dirección Meteorológica de Chile, 2016), which is
239 located outside the urban area, about 3 kilometers south of the downtown area close to
240 a former aerodrome.

241

242 2.3 Statistical analysis and imputation methods

243 Hourly air pollution data was converted to daily means according to the national
244 legislation (Ministerio del Medio Ambiente, 2018). After an initial analysis of
245 completeness, the missing data mechanism was diagnosed using two tests: Little's
246 MCAR test (Little, 1988) and the test of missingness (Schafer and Graham, 2002). The

247 data distribution was explored for all variables through histograms, Q-Q plots and the
 248 Shapiro-Wilk to test normality (Figure S 1). As distributions of PM_{2.5} and PM₁₀ were
 249 heavily skewed, they were log-transformed, which improved their performance and were
 250 used in further analysis. Descriptive analyses were performed for all variables including
 251 mean, median, percentiles and measures of dispersion (Table S1- S 2), along with
 252 boxplots by year (Figure S2), season (Figure S3) and precipitations (Figure S4). To
 253 explore associations between variables, bivariate analyses were performed, including
 254 scatterplot and Pearson correlations for continuous variables and boxplots, t-test and
 255 one-way ANOVA, for categorical ones.

256
 257 To reconstruct the datasets, five imputations methods were used: mean imputation,
 258 conditional mean imputation, K-Nearest Neighbor imputation, multiple imputation and
 259 Bayesian Principal Component Imputation, which are all based on multivariate
 260 regression models of PM_{2.5}. We built two initial regression models using log-transformed
 261 PM_{2.5} and usual covariates, as previously done (Díaz-Robles et al., 2008; Koutrakis et
 262 al., 2005; Sax et al., 2007). Model 1 included meteorological and temporal covariates, as
 263 well as PM₁₀ from the same monitoring station, as shown in Equation 1.

264

$$\begin{aligned}
 \ln(PM_{2.5}) = & \alpha + \sum \beta_{pm} * p_i + \sum \beta_t * t_i + \sum \beta_w * w_i + \sum \beta_{rh} * rh_i + \\
 & + \sum \beta_p * p_i + \sum \beta_y * y_i + \sum \beta_m * m_i + \sum \beta_d * d_i + \sum \beta_h * h_i + \varepsilon_i
 \end{aligned}
 \tag{Equation 1}$$

265
 266
 267 Where, α is the regression intercept; β_{pm} , β_t , β_w , β_{rh} , β_p , β_y , β_m , β_d , and β_h are the
 268 regression coefficients of the independent variables: $\ln(PM_{10})$, pm_i ; mean temperature, t_i ;

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

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

279 $f_i = \exp^{\beta x_i}$ Equation 2

280 $PM_{2.5} = \alpha \cdot f_{p,i} \cdot f_{t,i} \cdot f_{w,i} \cdot f_{rh,i} \cdot f_{p,i} \cdot f_{y,i} \cdot f_{m,i} \cdot f_{d,i} \cdot f_{h,i}$ Equation 3

281 With f_i being the concentration impact factor for any given regression estimate β for
 282 variable x in observation i ; α being the $PM_{2.5}$ concentrations when all covariates hold
 283 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
 284 impact factors for $\ln(PM_{10})$, temperature, relative humidity, precipitations, year, month,
 285 day of the week and holiday, respectively. Notice that a sensitivity analysis was
 286 performed using Reduced Major Axis (RMA) regression to examine the functional
 287 relationship between $PM_{2.5}$ and PM_{10} .

288
 289 Subsequently, the five imputations methods were applied to reconstruct the dataset. The
 290 first method was single imputation using the mean, where missing $PM_{2.5}$ values were
 291 replaced by the mean. The second imputation method, i.e., single imputation using

292 conditional mean, where missing $PM_{2.5}$ values were replaced by estimates from the
293 multiple linear regression model for all observations with complete covariates data. The
294 third method was K-Nearest Neighbor imputation. Here, we used the "mi impute pmm"
295 command in STATA 13 (StataCorp, College Station, TX) with 20 imputation sets and the
296 10 nearest neighbors. The command fills in the missing data with the closest values
297 based on the absolute difference between the linear prediction for the missing value and
298 the complete values. The fourth method was multiple imputation and was carried out
299 using the 'mi' command in STATA 13 (StataCorp, College Station, TX). Basically,
300 multiple imputation works through two stages—the imputation stage and the analysis
301 stage. The imputation stage creates imputations through an iterative Markov Chain
302 Monte Carlo process, assuming a multivariate normal underlying model. Twenty
303 imputations were executed, and each imputation iterated 2000 times, generating
304 complete datasets for both predictors and covariates. The convergence of the algorithm
305 was verified by examining autocorrelation and trace plots of imputed values. Each
306 completed dataset was verified to determine if the imputation process was complete. In
307 the analysis stage, final model parameters were estimated by combining each result
308 using Rubin's combination rules (StataCorp.Ltd, 2013). Finally, Bayesian Principal
309 Component imputation was employed. The number of principal components for each
310 model was selected. Then, an Expectation–maximization approach along with a
311 Bayesian model was employed to calculate the likelihood for a reconstructed value
312 (Stacklies et al., 2007) .

313 2.4 Validation datasets and evaluation of model performance

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

324
325 To evaluate the performance of environmental models, each imputed quarter was
326 compared against the original set separately, using five indicators commonly used to
327 assess the performance of environmental models (Bennett et al., 2013): i) Coefficient of
328 determination (R^2), ii) Root of the mean square error (RMSE), iii) Mean Absolute Error
329 (MAE), iv) Index of Agreement (IA), and v) Bias (B), as described in Equations 4-8:

330
$$R^2 = \left(\frac{\sum_{i=1}^n (y_i - \bar{y})(\hat{y}_i - \bar{\tilde{y}})}{\sqrt{\sum_{i=1}^n (y_i - \bar{y})^2} \sqrt{\sum_{i=1}^n (\hat{y}_i - \bar{\tilde{y}})^2}} \right)^2$$
 Equation 4

331
$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}$$
 Equation 5

332
$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|$$
 Equation 6

333
$$IA = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (|\hat{y}_i - \bar{y}| + |y_i + \bar{y}|)^2}$$
 Equation 7

334
$$B = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)$$
 Equation 8

335

336 Where, y_i and \hat{y}_i are the i th observation for the reconstructed and the comparison
 337 datasets, while \bar{y} and \tilde{y} are the means for the reconstructed and comparison datasets.
 338 R^2 is a squared version of the Pearson correlation coefficient and ranges from 0 (bad) to
 339 1 (good). It indicates how well the model explains the variance in the observations,
 340 compared with using their mean as the prediction. RMSE expresses the error in a metric
 341 that is in the same units as the original data. MAE is similar to RMSE except that the
 342 absolute value is used instead, thus, reducing the bias towards large events. IA, in turn,
 343 resembles to the coefficient of determination but is designed to better handle differences
 344 in modeled and observed means and variances. Finally, B calculates the mean error and
 345 indicates if the model tends to under- or over-estimate the measured data with an ideal
 346 value of zero. For log-transformed variables, the exponential form informs us the relative
 347 error or bias, and can be expressed as percentage (%).

348 3 Results

349 3.1 Data completeness and pattern of missingness.

350 Table 1 shows data completeness for the monitoring stations. In general, completeness
351 of PM₁₀ and PM_{2.5} was not very high, with losses of the order of 20%, and a slightly
352 better performance of *Las Encinas* compared to *Museo Ferroviario*. For the other
353 pollutants (NO_x, CO, O₃), completeness was even worse. This highlights the need to
354 reconstruct the PM datasets, as a large portion of the health data would not have
355 exposure data available. Meteorological variables presented a much better performance,
356 particularly at the *Maquehue* station, so it was used for the regression models.

357
358 The pattern of missingness is shown in Table 2. When considering PM₁₀, PM_{2.5}, and
359 meteorological variables (temperature, relative humidity, precipitation, and wind speed)
360 at *Las Encinas*, the main pattern is complete case (76%), followed by missing PM_{2.5} and
361 PM₁₀ (9%), and PM_{2.5} only (7%) with all other patterns being negligible. A similar pattern
362 is observed for the *Museo Ferroviario* dataset. The Little test obtained a Chi² of 762 (df:
363 72, p<0.01), indicating that the data seems to be MAR because there exists an
364 identifiable pattern for the missing data. In addition, the test of missingness for
365 independence showed that data was MAR with losses associated with other variables in
366 the dataset: PM₁₀ (OR=1.5; p<0.01), years (overall p<0.01), March (OR=0.3; p<0.01),
367 April (OR=0.4; p<0.01), September (OR=0.5; p=0.05), and October (OR=0.5; p=0.02)
368 (Table S4).

369

370 Table 1. Data completeness for Temuco air quality and meteorological stations.
 371

Year	Pollutants										Meteorological variables											
	PM _{2.5}		PM ₁₀		NO _x		O ₃		CO		Temperature			RH			Wind speed			Precipitation		
	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

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

373 *Wind speed: scalar average

374

375 Table 2. Missing data patterns for the *Las Encinas*, *Museo Ferroviario* and *Maquehue*
 376 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	

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

378

379

380

381 3.2 Variable characterization

382 Table 3 and Figure S2 show summary statistics and distributions for PM_{2.5} and PM₁₀.
 383 Overall, PM_{2.5} and PM₁₀ concentrations exceeded national standards and international
 384 guidelines with PM_{2.5} concentrations being significantly above the national annual
 385 standard of 20 µg/m³ (Ministerio de Medio Ambiente, 2014) and the WHO annual Air
 386 Quality Guideline of 10 µg/m³ (World Health Organization, 2006). Many days exceeded
 387 the national daily standard of 50 µg/m³, and even reached concentrations as high as 200
 388 µg/m³. PM₁₀ also showed concentrations above standards, but mainly driven by PM_{2.5},
 389 as about 80% of PM₁₀ is comprised of PM_{2.5} (Ministerio del Medio Ambiente, 2015).

390

391 Table 3. Summary statistics for PM_{2.5} and PM₁₀, by year and station.

Year	PM _{2.5}				PM ₁₀			
	Las Encinas		Museo Ferroviario		Las Encinas		Museo Ferroviario	
	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

392 SD: standard deviation.

393

394 The bivariate analyses (Figure 2) show strong associations of PM_{2.5} with temporal
 395 variables such as some years (with no temporal trend) and month (higher in winter), but
 396 not with weekday or weekends. Additional associations were observed with PM₁₀
 397 (directly associated), temperature (higher when cold), relative humidity (higher when
 398 humid), and wind speed (higher when stagnant), but not with precipitations. When
 399 analyzing hourly patterns (Figure S3), highest concentrations were observed at night

400 from 6 pm to 4 am, independent of the day of the week, with the pattern more
401 pronounced in winter, and with little evidence of other peaks associated with traffic-rush
402 hours. These patterns are all in agreement with small, residential wood-burning particles
403 being the main source of $PM_{2.5}$, which persist in summer due to the use of stoves for
404 cooking, although attenuated.

405

406 3.3 Regression model and imputation.

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

423 sensitivity analysis of $PM_{2.5}$ and PM_{10} in both models with the RMA regression (Table S
424 5).

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

Effect	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			
	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

426 The estimates are expressed as one-unit increase in the predictor. Reference variables are 2009, January, Sunday
427 and working day. *Overall p-value for the variable. CIF: concentration impact factor. CIF is referred to changes in
428 predictors of: $\Delta PM_{10}=10\%$; $\Delta PM_{2.5}=10\%$; $\Delta Temp=5^{\circ}C$; $\Delta WS=10knots$; $\Delta RH=10\%$; na= not applicable; Wind speed:
429 scalar average
430
431

432 3.4 Performance of imputation methods on validation datasets.

433 The results of the imputation methods on full and validation datasets are shown in Table
434 5, Figures S 5- S 6. In general, K-Nearest Neighbor presented a better performance
435 than other imputations methods in both full and validation datasets. However, to the
436 contrary of multiple imputation, K-Nearest neighbor was unable to reconstruct the full
437 dataset because of missing values in the covariates (keeping missing data about 12%)
438 (Figure S5). Model performance improved when including data from another station
439 (*Museo Ferroviario*, Model 2) (Figure 3). For the full dataset, multiple imputation using
440 model 2 provided the highest completeness (100%) with a lower error ($e^{RMSE}=27\%$,
441 $e^{MAE}=24\%$), and lower bias ($e^{Bias}=3.9\%$), thus being a promising option to reconstruct the
442 Temuco dataset. The lower performance was observed for Bayesian principal
443 component imputation for both models. When challenged with the validation datasets,
444 the performance remained for most indicators and most datasets, but decreased slightly
445 for R^2 and IA, in general, and particularly for some sets. In addition, for some sets (p25 -
446 p75), bias was away from 0 on the order of 10%-20%, indicating that in some cases a
447 small bias can be introduced in the set due to the imputation process.

448 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 (p25-p75)	Median (p25-p75)	Median (p25-p75)	Median (p25-p75)	Median (p25-p75)	Median (p25-p75)
Model 1:						
Mean Imputation	80 (63-88)	0.80 (0.46-0.90)	26 (19-28)	28 (24-34)	2.9 (-7.4-16.4)	0.92 (0.76-0.96)
Conditional Mean Imputation	80 (63-88)	0.80 (0.45-0.89)	27 (21-30)	28 (22-32)	4.3 (-12.5-9.8)	0.91 (0.78-0.97)
K-Nearest Neighbor	80 (63-88)	0.80 (0.45-0.89)	28 (21-30)	28 (22-32)	4.1 (-12.1-9.6)	0.90 (0.78-0.97)
Multiple Imputation	82.5 (66-10)	0.78 (0.41-0.89)	29 (21-33)	33 (27-43)	7.7 (-21.9-17.4)	0.87 (0.72-0.95)
Bayesian Principal component analysis	82 (65-89)	0.75 (0.37-0.84)	29 (21-31)	40 (30-54)	9.2 (-19.9-32.0)	0.89 (0.62-0.92)
Model 2:						
Mean Imputation	71.5 (25-82)	0.83 (0.73-0.91)	20 (18-24)	22 (18-26)	0.6 (-7.8-6.5)	0.95 (0.92-0.97)
Conditional Mean Imputation	72 (25-82)	0.85 (0.74-0.91)	21 (19-26)	22 (19-27)	-1.8 (-7.8-5.6)	0.95 (0.91-0.97)
K-Nearest Neighbor	80 (63-88)	0.80 (0.45-0.89)	28 (21-30)	28 (22-32)	4.1 (-12.1-9.6)	0.90 (0.78-0.97)
Multiple Imputation	83 (66-90)	0.81 (0.61-0.90)	26 (20-31)	31 (22-37)	-2.8 (-2.8--13.8)	0.92 (0.81-0.96)
Bayesian Principal component analysis	82 (65-89)	0.79 (0.55-0.86)	25 (20-31)	37 (24-51)	5.2 (-19.9-24.7)	0.89 (0.69-0.94)

449 Obs: Observations; RMSE: Root mean square error; MAE: Mean absolute error, IA: Index of agreement .

450 *RMSE(%)=[exp(RMSE)-1]*100; **MAE(%)=[exp(MAE)-1]*100; ***Bias(%)=[exp(Bias)-1]*100

451 **4 Discussion**

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

460

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

468

469 Rather few previous studies have used imputation methods to reconstruct datasets. In a
470 comprehensive study using data with missingness near 25% from Helsinki, Finland,
471 and Belfast, North Ireland; similar measures of performance were found with R^2 of 0.49,
472 RMSE of 0.22 and MAE of 0.16 (Junninen et al., 2004). Additionally, they found that
473 single imputation methods underestimated the error variance and accuracy of missing

474 data compared to multiple imputation, which might explain our results. In another study
475 using datasets in La Coruña, Spain, several imputation methods were compared
476 (Gómez-Carracedo et al., 2014). They used factor analysis with Varimax rotation along
477 with the imputation methods, but did not provide overall performance measures, in terms
478 of completeness, error, and bias, and did not challenge the methods with validation sets.
479 They found that multiple imputation had more scattered results when datasets had more
480 than 43.5% of missingness and were poorly correlated with other variables; however,
481 results were similar when missingness was medium, as in our case. Finally, an infant
482 cohort study investigating the effects of pollution on asthma risk (Roda et al., 2014),
483 compared methods for imputing indoor domestic pollutants. The complete case reduced
484 the statistical power, while single imputation overestimated the association and multiple
485 imputation was too conservative and unable to show significant associations.

486 Considering this experience, it seems necessary that researchers continue attempting
487 the reconstruction of datasets, particularly where more needed, such as low- middle-
488 income countries and small cities. It seems important to provide overall indicators of
489 performance, as these can be locally driven by the quality of the data and the base
490 regression model. Junger and de Leon (2015) developed a time-series for an air
491 pollution simulation study using complete case analysis, unconditional mean imputation,
492 conditional mean imputation and other approaches such as a regular Expectation
493 Maximization algorithm (EM), EM algorithm filtered by splines, among others. They
494 found that when the amount of missing data was less than 5%, the complete case
495 analysis had a good performance. However, when the missing data was higher the
496 validity of estimates degraded.

497

498 The results are limited only to Temuco and for the time-period under study. The
499 combination of explanatory variables selected in our imputation models for Temuco
500 might differ in other locations. For instance, the application of this framework to areas
501 located near large industrial complexes or surface mining operation might highlight wind
502 direction to be a strong predictor for ambient $PM_{2.5}$, whereas the model for Temuco did
503 not include this variable in the final model. Similarly, cities located in arid regions have a
504 larger influence from coarse particles, weakening the correlation between PM_{10} and
505 $PM_{2.5}$. However, the methodological framework employed in this study to identify the
506 best imputation model could be usefully replicated in other regions and cities. Therefore,
507 it would be interesting to extend the current approach to other time periods in Temuco,
508 other cities in Chile and elsewhere, taking into in consideration the specific atmospheric
509 composition, sources and dynamics of the air shed in individual cities.

510
511 A limitation of this work is the fact that the background concentration of air pollution or
512 the boundary layer are not measured by the monitoring air quality network and could not
513 be included in the statistical models. However, previous research in the study area have
514 shown that the main source of air pollution is residential wood burning (Jorquera et al.,
515 2018; SICAM, 2014; Villalobos et al., 2017, 2015). A potential limitation of using
516 imputation methods to predict missing values would occur in the case that the data were
517 MNAR, as it might introduce bias in the data set. Results from our validation dataset,
518 showed small bias in general, but more significant in some specific cases like Bayesian
519 principal component analysis. This is a warning as in some circumstances a bias in
520 $PM_{2.5}$ estimation might be introduced even if the MAR assumptions would be met;
521 however, this bias seems not to be high, on the order of 10%-20%. In any circumstance,

522 the possibility of biasing the health estimates due to the introduction of a small bias
523 during the imputation process should be weighed against the possible bias incurred by
524 not including the full dataset in the analysis.

525
526 In summary, our results show that using imputation methods, particularly multiple
527 imputation, can be to a certain extent successful in reconstructing an air quality data set
528 with relatively low-medium missingness in a real-life situation. This is relevant for
529 datasets in small locations where the problem of missing data might be more frequent
530 alongside with serious environmental health problems.

531

532

533

534

535 **Acknowledgement**

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

545

546 **Competing financial interests**

547 The authors declare they have no competing interests.

548 **5 References**

- 549 Bennett, N.D., Croke, B.F.W.W., Guariso, G., Guillaume, J.H.A.A., Hamilton, S.H., Jakeman,
550 A.J., Marsili-Libelli, S., Newham, L.T.H.H., Norton, J.P., Perrin, C., Pierce, S.A., Robson, B.,
551 Seppelt, R., Voinov, A.A., Fath, B.D., Andreassian, V., 2013. Characterising performance of
552 environmental models. *Environ. Model. Softw.* 40, 1–20.
553 <https://doi.org/10.1016/j.envsoft.2012.09.011>
- 554 Bishop, C.M., 1999. Variational principal components. *IEE Conf. Publ. Artif. Neural Networks*
555 509–514.
- 556 Díaz-Robles, L.A., Ortega, J.C., Fu, J.S., Reed, G.D., Chow, J.C., Watson, J.G., Moncada-
557 Herrera, J.A., 2008. A hybrid ARIMA and artificial neural networks model to forecast
558 particulate matter in urban areas: The case of Temuco, Chile. *Atmos. Environ.* 42, 8331–
559 8340. <https://doi.org/10.1016/j.atmosenv.2008.07.020>
- 560 Dirección Meteorológica de Chile, 2016. Climatología. Available from
561 <http://www.meteochile.cl/PortalDMC-web/index.xhtml>.
- 562 Dixon, J.K., 1979. Pattern recognition with partly missing data. *IEEE Trans. Syst. Man, Cybern.*
563 10 617–621.
- 564 Gómez-Carracedo, M.P., Andrade, J.M., López-Mahía, P., Muniategui, S., Prada, D., 2014. A
565 practical comparison of single and multiple imputation methods to handle complex missing
566 data in air quality datasets. *Chemom. Intell. Lab. Syst.* 134, 23–33.
567 <https://doi.org/10.1016/j.chemolab.2014.02.007>
- 568 Gómez, W., Salgado, H., Vásquez, F., Chávez, C., 2014. Using stated preference methods to
569 design cost-effective subsidy programs to induce technology adoption: An application to a
570 stove program in southern Chile. *J. Environ. Manage.* 132, 346–357.
571 <https://doi.org/10.1016/j.jenvman.2013.11.020>
- 572 Green, J., Sánchez, S., 2012. La Calidad del Aire en América Latina: Una Visión Panorámica.
573 Clean Air Institute. Available from
574 http://www.minambiente.gov.co/images/AsuntosambientalesySectorialyUrbana/pdf/contaminacion_atmosferica/La_Calidad_del_Aire_en_Am%C3%A9rica_Latina.pdf.
- 576 Greenland, S., Finkle, W.D., 1995. A critical look at methods for handling missing covariates in
577 epidemiologic regression analyses. *Am. J. Epidemiol.* 142, 1255–1264.
578 <https://doi.org/https://doi.org/10.1093/oxfordjournals.aje.a117592>
- 579 INE, 2017. Anuarios parque de vehículos en circulación. Available from
580 http://historico.ine.cl/canales/chile_estadistico/estadisticas_economicas/transporte_y_comunicaciones/parquevehiculos.php.
- 582 James, G., Witten, D., Hastie, T., Tibshirani, R., 2015. Resampling methods, in: *An Introduction*
583 *to Statistical Learning*. pp. 176–184.
- 584 Jorquera, H., Barraza, F., Heyer, J., Valdivia, G., Schiappacasse, L.N., Montoya, L.D., 2018.
585 Indoor PM_{2.5} in an urban zone with heavy wood smoke pollution: The case of Temuco,
586 Chile. *Environ. Pollut.* 236, 477–487. <https://doi.org/10.1016/j.envpol.2018.01.085>
- 587 Junger, W., de Leon, A.P., 2009. Missing Data Imputation in Time Series of Air Pollution.
588 *Epidemiology* 20. <https://doi.org/10.1097/01.ede.0000362970.08869.60>
- 589 Junger, W.L., de Leon, A.P., 2015. Imputation of missing data in time series for air pollutants.

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

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

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

597 Koutrakis, P., Sax, S.N., Sarnat, J. a, Coull, B., Demokritou, P., Oyola, P., Garcia, J., Gramsch,
598 E., 2005. Analysis of PM₁₀, PM_{2.5}, and PM_{2.5-10} concentrations in Santiago, Chile, from
599 1989 to 2001. J. Air Waste Manag. Assoc. 55, 342–351.
600 <https://doi.org/10.1080/10473289.2005.10464627>

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

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

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

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

611 Ministerio de Medio Ambiente, 2014. Planes de Descontaminación Atmosférica Estrategia 2014
612 - 2018. Available from [http://portal.mma.gob.cl/planes-de-descontaminacion-atmosferica-](http://portal.mma.gob.cl/planes-de-descontaminacion-atmosferica-estrategia-2014-2018/)
613 [estrategia-2014-2018/](http://portal.mma.gob.cl/planes-de-descontaminacion-atmosferica-estrategia-2014-2018/).

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

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

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

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

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

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

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

634 <https://doi.org/10.1371/journal.pone.0104254>

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

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

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

642 Sax, S.N., Koutrakis, P., Ruiz Rudolph, P.A., Cereceda-Balic, F., Gramsch, E., Oyola, P., 2007.
643 Trends in the elemental composition of fine particulate matter in Santiago, Chile, from 1998
644 to 2003. J. Air Waste Manag. Assoc. 57, 845–855. [https://doi.org/10.3155/1047-
645 3289.57.7.845](https://doi.org/10.3155/1047-3289.57.7.845)

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

648 SICAM, 2014. Emission Inventory for the Temuco-Padre Las Casas Metropolitan Area: Year
649 2013: Residential Wood Burning. Temuco.

650 Stacklies, W., Redestig, H., Scholz, M., Walther, D., Selbig, J., 2007. pcaMethods – a
651 Bioconductor package providing PCA methods for incomplete data. Bioinformatics 23,
652 1164–1167.

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

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

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

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

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

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

671 Villalobos, A.M., Barraza, F., Jorquera, H., Schauer, J.J., 2015. Chemical speciation and source
672 apportionment of fine particulate matter in Santiago, Chile, 2013. Sci. Total Environ. 512–
673 513, 133–142. <https://doi.org/10.1016/j.scitotenv.2015.01.006>

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

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

679

680

Figure 1

Great Temuco



Temuco

Museo Ferroviario

Las Encinas

Cautin River

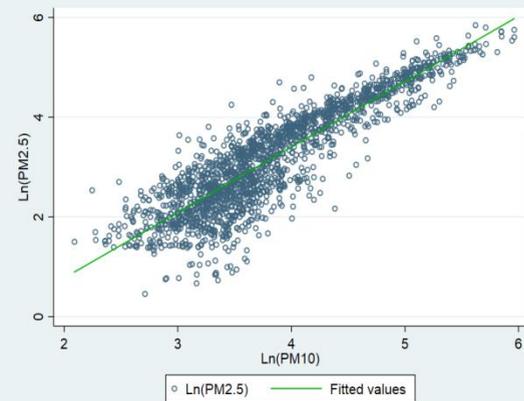
Padre Las Casas

Maquehue

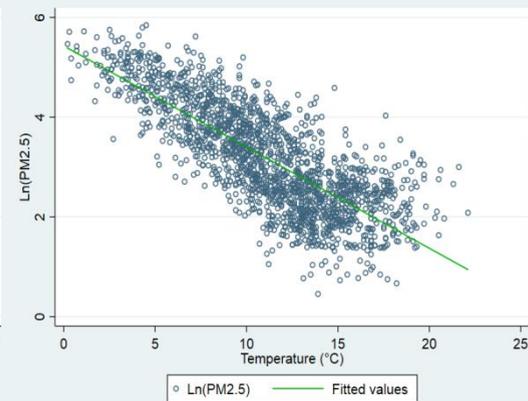
Padre Las Casas



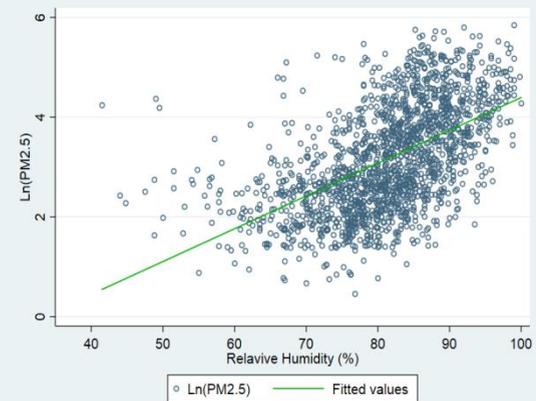
Figure 12 ($R^2=0.79$, $p<0.001$)



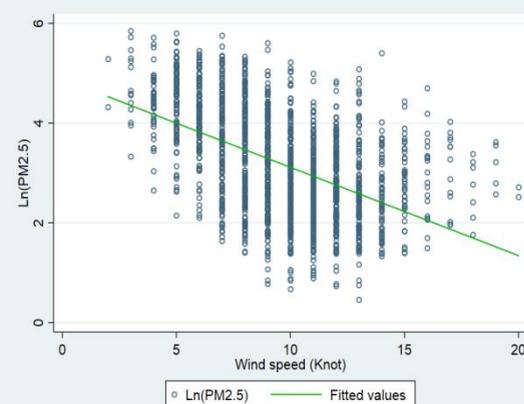
b) Temperature ($R^2=0.60$, $p<0.001$)



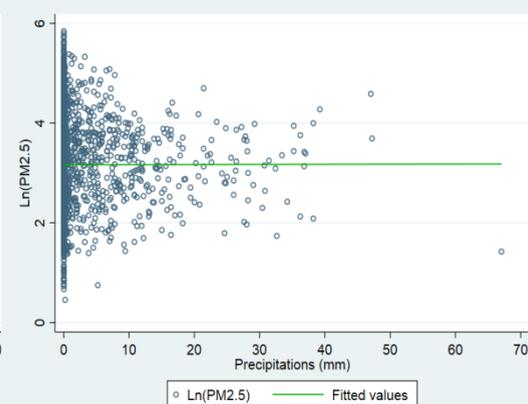
c) Relative humidity ($R^2=0.30$, $p<0.001$)



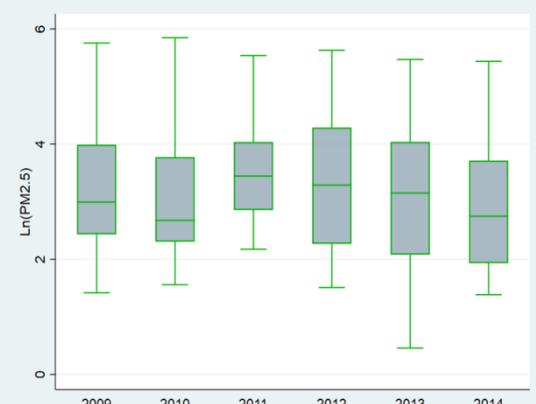
d) Wind speed ($R^2=0.25$, $p<0.001$)



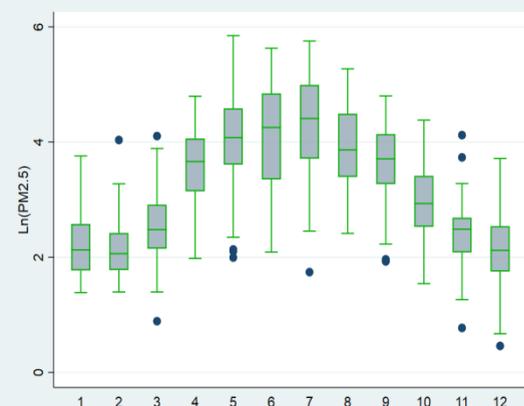
e) Precipitation ($R^2=0.07$, $p<0.001$)



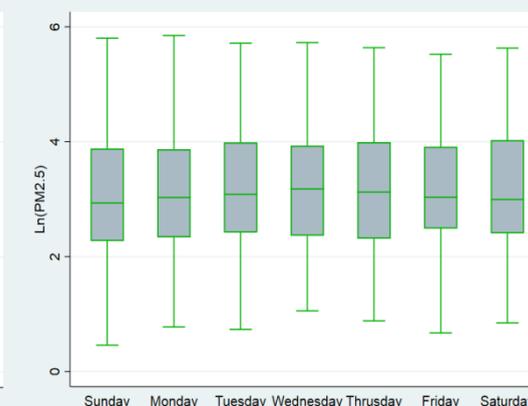
f) Year ($F=17.85$, $p<0.001$)



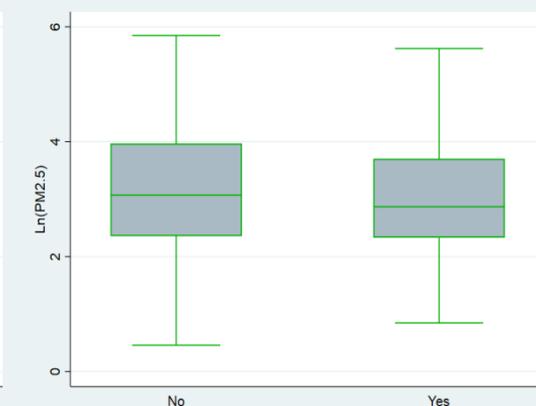
g) Month ($F=261.95$, $p<0.01$)



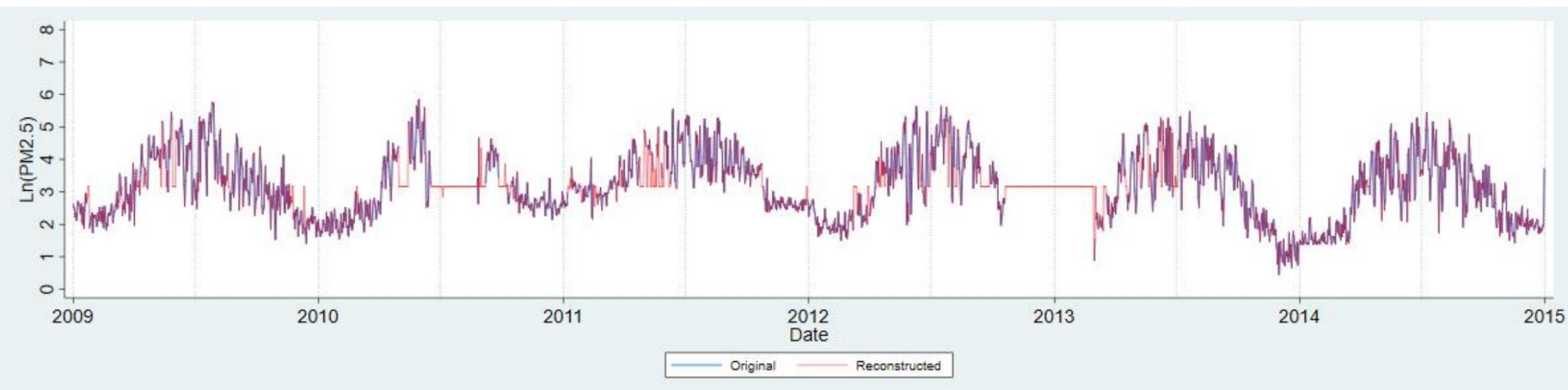
h) Day of the week ($F=0.29$, $p=0.96$)



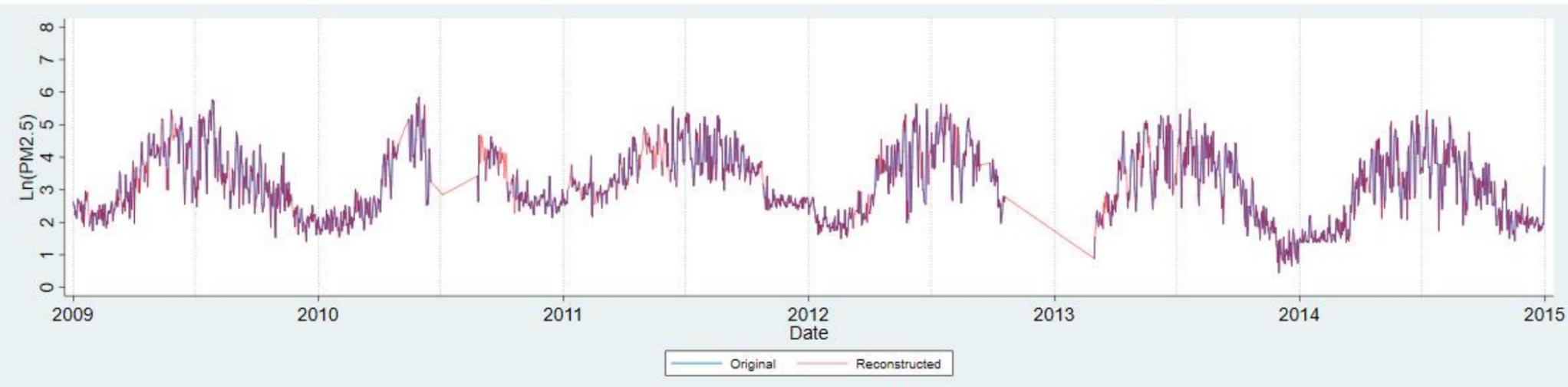
i) Holiday ($F=2.49$, $p=0.11$)



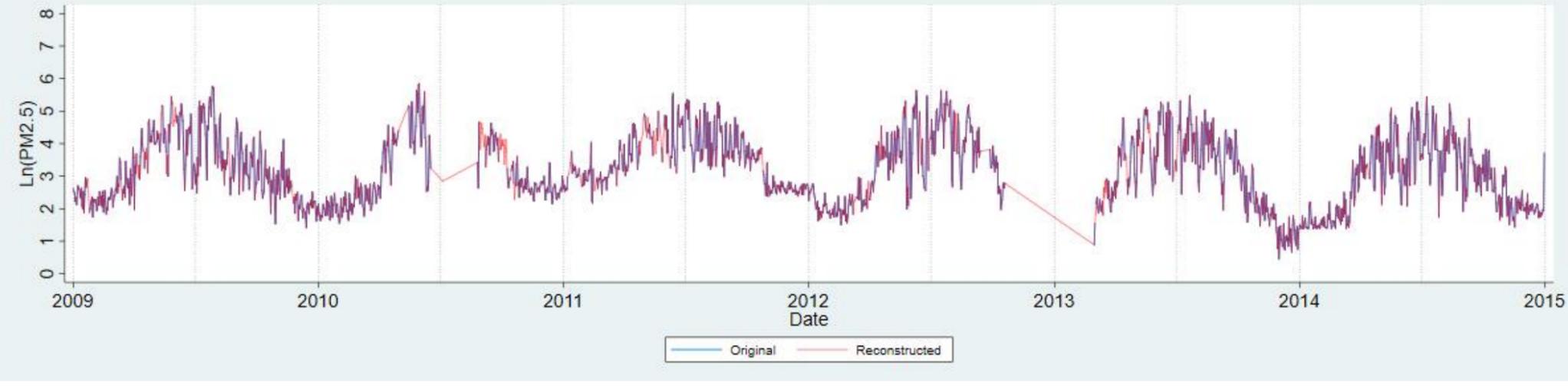
a) Mean Imputation



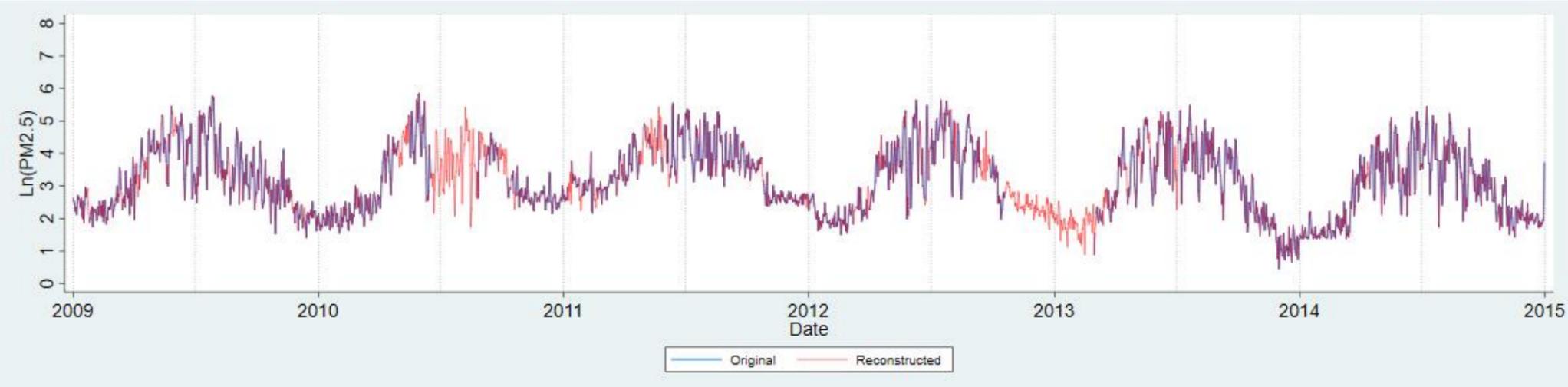
b) Conditional Mean Imputation



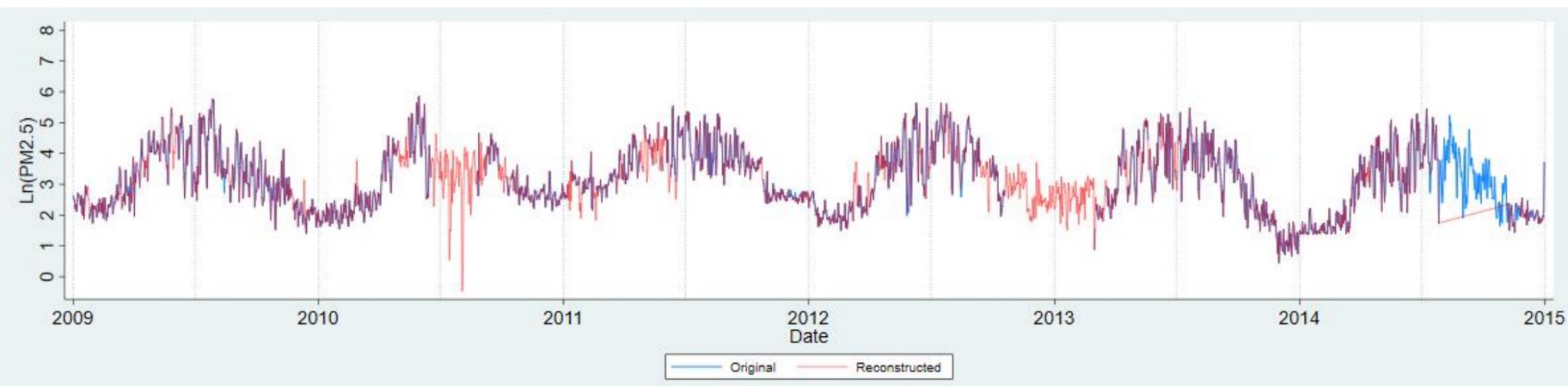
c) K Nearest Neighbord Imputation



d) Multiple Imputation



e) Bayesian Principal Analysis Imputation



Supplementary Material

[Click here to download Supplementary Material: Quinteros ME_MI SUPP 181109.pdf](#)