Instituto Tecnológico y de Estudios Superiores de Monterrey

Campus Monterrey

School of Engineering and Sciences

Complementary inverse modeling and machine learning techniques for air pollution assessment

A dissertation presented by

Ana Yael Vanoye García

Submitted to the
School of Engineering and Sciences
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

In

Engineering Science

Major in Environmental Systems

Monterrey Nuevo León, December 5th, 2022

To my father Arnoldo Vanoye and my mother Rebeca García
with all my love and gratitude.

To my sister Patsy and my brother Arnoldo,
you are angels to me.

Acknowledgments

My deepest gratitude to Dr. Alberto Mendoza Domínguez, for making this project possible, from the very beginning to the end. Thank you for all the opportunities and learning experiences, and for sharing all your knowledge with us and being a role example. In your research group, I found a purpose, vocation, and even a second family.

To Isabel Cristina Xicoténcatl, Diego Alejandro González and Sergio Santiago Cárdenas, for your commitment and hard work on this project, and for everything I learned with you.

To my Thesis Committee: Dra. Fabiola Yépez, Dr. Alejandro Álvarez, Dr. Gerardo Mejía and Dr. José I. Huertas, for your patient advice and recommendations, that I will take seriously and surely will transcend this work.

To the many wonderful friends and allies I have met on the way. Special thanks to Natalia, Lizzy, Delia, Vicky, Johana, Mónica R., Mónica M., Maritza, Aida, Jessica, Belén, Hugo, Eder, Alejandro, Edson, Gerardo, for your continuous encouragement and support, for being with me through thick and thin.

To my dear colleagues from the Department of Sustainable Technologies: Mary, Gloria, Gaby, Irma, Mónica, Magda, Emma, Sadot, Carlos W., Carlos H., for your support and solidarity.

To my academic and work leaders, for providing opportunities and spaces for my professional development.

To my students, for cheering me and inspiring me to be a better professor.

To Dalia, Nydia, Lolita and Gaby, for being home to me.

Finally, I want to thank CONACYT and Tecnológico de Monterrey, for assisting with funds and scholarship to complete this research.

Abstract

Ambient air pollution is considered the greatest environmental threat to human health. Air quality models (AQM) describe the atmospheric dynamics of air pollutants and can help identify source contributions to air quality. Deterministic AQM estimate the relationship between sources of pollution and their effects on ambient air quality by simulating the evolution over time of three-dimensional fields of concentrations of pollutant species. However, a key input to deterministic AQM are detailed, spatially and temporally-resolved emission inventories, which are known to carry large uncertainties. On the other hand, recent advances in data science allows for the use of supervised machine learning methods, such as Multivariate Linear Regression Model (MLRM) along socioeconomic historical data to explore the evolution of pollution sources through time. In this work, the use of inverse modeling mathematical tools to improve emission inventories for deterministic air quality modeling applications, and the application of supervised machine learning tools to quantify the contribution of energy and economic factors to air pollution are explored. An analysis performed on emission inventories developed for the Monterrey Metropolitan area exhibited large differences (in the range of -3.6% to +51.7%) for recent published criteria pollutant inventories and demonstrated the need to assess techniques to reduce emission inventories uncertainty. A literature review on regularization methods showed that these mathematical techniques are increasingly being used in the atmospheric sciences in inverse modeling contexts. As study case, some regularization methods (namely, Tikhonov regularization, Truncated Singular Value Decomposition, and Damped Singular Value Decomposition) in combination with regularization parameter selection methods (Generalized Cross Validation, L-Curve, and Normalized Cumulative Periodograms), along a Bounded Variable Least Squares method, were used with a deterministic photochemical air quality model to compute scaling factors for the correction of a criteria-pollutant emission inventory for Guadalajara Metropolitan Area, Mexico. The inverse modeling with regularization approach was able to adequately resolve ozone concentrations, a secondary pollutant, by adjusting its precursor emissions, obtaining Daily Indices of Agreement up to 0.95 (compared to 0.89 of the base case). Also, the non-systematic error was reduced. However, results also reflected that regularization methods alone cannot resolve all uncertainties, and that incorporating known available data could be useful to better understand pollution sources. Therefore, MLRM were developed by correlating long-term economic and energy indicators, routinely reported by Mexican government agencies, with monthly-averaged air pollution data for the Monterrey, Guadalajara and Mexico City Metropolitan Areas. Although socioeconomic variables do not explain all variance in the observed pollutant concentrations, they allow the identification and analysis of activities with impact in air quality. Moreover, the resulting MLRM models displayed similar statistical performance and compared favorably to other studies found in literature. It is concluded that both approaches (deterministic and machine-learning models) are data-driven at core and can help design relevant public policies.

This page intentionally left blank

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

Air quality modeling estimates the relationship between sources of pollution and their effects on ambient air quality. Air quality models describe the atmospheric dynamics of air pollutants, can help identify source contributions to air quality, and one of their main applications is in the development of strategies to reduce their concentration (Environmental Protection Agency, 2022). In the last decades, a variety of different air quality models have been developed, ranging from those based upon simple to the most sophisticated approaches. Because models differ one of another, different model attributes can be used for their classification. Table 1.1 lists some of the most important types of air quality models. The most significant factor, however, which divides air pollution models into two independent groups is the basic model structure: deterministic and non-deterministic models (Juda-Rezler, 1991).

Table 1.1 Types of air quality models based on their attributes (Juda-Rezler, 1991)

Attribute		**Size**	**Time horizon**	**Pollutant of concern**

Basic model structure	• Deterministic
	• Non-deterministic
	  ○ Statistical
	  ○ Physical (fluid)
		• Local
		• Regional
		• National
		• Global
			• Hour
			• Day
			• Month
			• Year
			• Decade
				• Criteria pollutants (e.g. SO_2, NO_x, HC, PM, CO)
				• Photochemical oxidants
				• Secondary pollutants
				• Greenhouse gases

Time resolution	• Steady-state • Time-dependent			
Frame of reference	• Eulerian • Lagrangian			
Dimensionality of computational domain	• 1D • 2D • 3D • Multilevel			
Methods of model equations resolution	• Analytical • Numerical (various methods)			

In general, deterministic air quality models attempt to simulate the evolution over time of three-dimensional fields of concentrations of pollutant species. The chemical species described by these models can be inert or reactive, and the models can be applied at local (close to emissions), regional or continental scales.

If c_i represents the concentration of the species i in gas phase, its evolution over time is governed by a differential equation of reaction-diffusion-advection:

$$\frac{\partial c_i}{\partial t} + \boldsymbol{div}\,(\boldsymbol{V c_i}) = \boldsymbol{div}\left(\boldsymbol{\rho K \nabla \frac{c_i}{\rho}}\right) + \boldsymbol{X_i(c,x,t)} - \boldsymbol{\Lambda_i(x,t)\,c_i} + \boldsymbol{S_i(x,t)} \qquad \textbf{(1.1)}$$

where c represents concentration, x stands for position, t is the time, V is the average wind speed, ρ is the density of the air, K is the matrix of turbulent dispersion coefficients, Λ_i is the drag coefficient, S_i is the term representing emissions by point sources (volumetric), as given by emission inventories, and X_i is the chemical source term that describes chemical reactions and whose response is nonlinear for many reactive species.

One of the most important components of air quality models is the photochemical mechanism that describes how volatile organic compounds (VOCs) and nitrogen oxides (NO_x) interact to produce ozone (O_3) and other oxidizing species (Yu et al., 2010). Thus, for air quality modeling applications, different mechanisms have been developed. One example is the chemical mechanism Carbon Bond (CB05) (Yarwood et al., 2005) which includes 59 species (between organic and inorganic) and 156 chemical reactions (Yu et al., 2010). Among the chemical species that CB05 can treat are ethane, olefins, terpenes, formic acid, acetic acid, methanol, ethanol, peroxyacetic acid, among others. Another alternative chemical mechanism is SAPRC99, which simulates the photochemistry of 80 species through 214 reactions. Although studies have been conducted comparing its performance in predicting specific type of pollutants, e.g., ozone, it has been concluded that no mechanism – among CB5, CB4 and SAPRC99 – performs systematically better than another (Yu et al., 2010).

Likewise, air quality models need to be fed with detailed meteorological information, particularly for parameters such as wind speed and direction, temperature, humidity, pressure and solar radiation. Typically, this information is generated using mesoscale meteorological models such as the Meteorological Mesoscale Model (MM5) (Grell et al., 1994) or the Weather Research Forecasting Model (WRF) (Skamarock et al. 2005). Despite the general uncertainties associated with the application of these models, they have been widely used and validated for air quality applications.

In addition to a robust chemical mechanism and reliable meteorological information, deterministic air quality models require detailed emission inventories, which even today carry great uncertainty despite the constant efforts made to improve them (Napelenok et al., 2011). Figure 1.1 (Li et al. 2021) presents a typical air pollution modeling system.

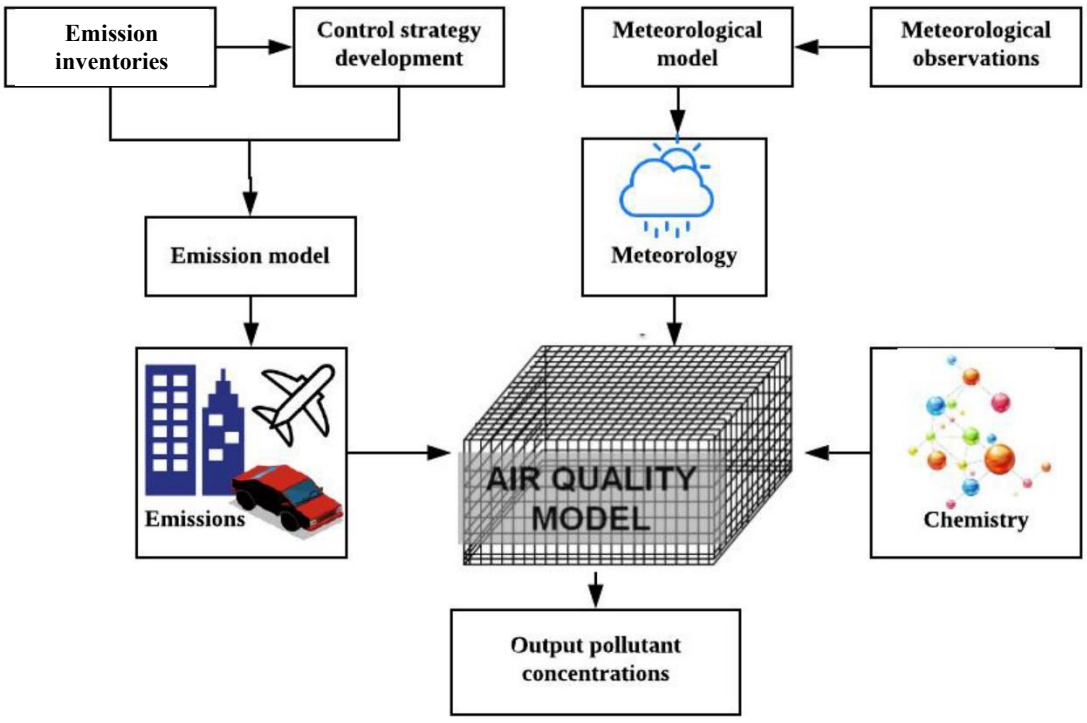

Figure 1.1. Air quality model and its inputs (From: Li et al. 2021).

On the other hand, non-deterministic models can be further divided in two groups: statistical and physical (fluid) models (Juda-Rezler, 1991). Statistical models calculate pollutants´ concentration by statistical methods from meteorological and other parameters after a statistical relationship has been obtained empirically from measured concentrations, while physical models are those in which nature is simulated on a smaller scale in the laboratory.

Most statistical models proposed in air pollution are applications of well-known statistical methods used in meteorology. They vary from simple contingency tables through univariate and multiple regression models to time-series techniques. The statistical models are essentially empirical, because even the most complex ones are based on a group of observations (Juda-Rezler, 1991).

Lately, the development of computer science and the continuous improvement and innovation of statistical prediction methods, has allowed the combination of traditional regression methods and spatial statistical methods into more complex analysis methods, such as machine learning (ML) approaches. ML refers to the automated detection of meaningful patterns in data, and strictly rely on historical data to make predictions. Machine learning algorithms can be classified according to the desired outcome of the algorithm. Supervised learning generates a function that maps inputs to desired outputs. The supervised learning task can be formulated as a classification problem: The learner is required to “learn”, in other words, to approximate the behavior of a function which maps a vector into one of several classes by looking at several input-output examples of the function (Osinsawo et al. 2017).

Some examples of supervised ML techniques include linear classifiers, logistic regression, Naïve Bayes classifier, perceptron, Support Vector Machine (SVM), quadratic classifiers, K-means clustering, boosting, decision trees, random forest (RF), neural networks (NN), and Bayesian networks (Osinsawo et al. 2017), which have emerged in the atmospheric sciences in recent years with promising results

(Carmona et al. 2021, Taheri Shahraiyni et al. 2016, Liao et al. 2021). Figure 1.2 presents a simplified representation of a typical supervised ML algorithm.

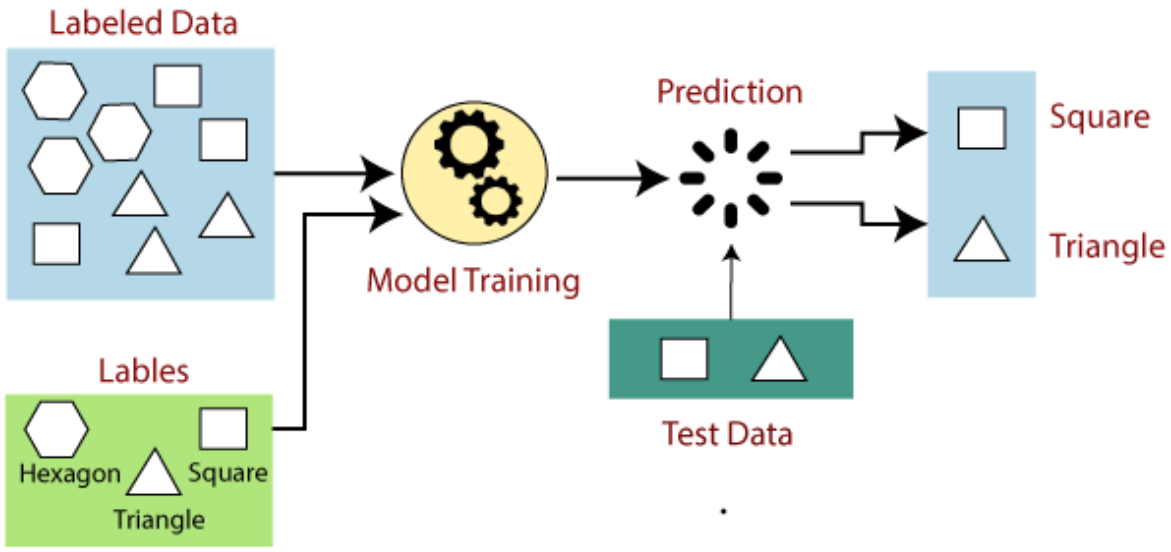

Figure 1.2. Simplified representation of a supervised machine learning algorithm. (From: Javatpoint, 2022).

Among supervised LM techniques, multivariate linear regression (MLR) models are one of the simplest algorithms. Regression analysis is a statistical tool that investigates relationships between variables. Usually, researchers seek to ascertain the causal effect of independent variables y upon dependent variables x_i. When a model is used to forecast $\boldsymbol{y}$ for a particular set of values of x_i, it is important to measure how large the error of the forecast might be. All these elements, including dependent and independent variables and error, are part of a regression analysis, and the resulting forecast equation is often called a regression model. Regression analysis is a basic technique in air pollution forecasting (Bai et al. 2018). Univariate linear regression can be expressed as:

$$\boldsymbol{y = b_0 + b_1 x + e} \qquad \textbf{(1.2)}$$

Multiple-linear regression (MLR) models are given as:

$$y = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_i x_i + e_i \quad (1.3)$$

where y is the dependent variable, x and x_i are the independent variables, b and b_i are the regression coefficients, and e is the error. It has a normal distribution with a mean of 0, which implies the conditional probability of y given x is normal too.

Wilks (2006) presents a thorough review of basic statistical tools and methods applied in atmospheric sciences, from empirical distributions and exploratory data analysis to statistical forecasting and multivariate statistics. Maçaira et al. (2018) performed a systematic literature review on time series analysis with explanatory variables, which encompasses methods to model and predict correlated data considering additional information. In the literature, multiple applications of MLR models to study air quality have been documented, e.g. Rosenlund et al. (2008) Ganesh et al. (2017), Bai et al. (2018), Ganesh et al. (2019), Abdullah et al. (2020), Shams et al. (2021), He et al. (2022).

Computational advances have also spurred the use of Computational Fluid Dynamics (CFD) models to study microscale pollutant dispersion (e.g. Pantusheva et al. 2022). CFD models can be classified as a type of deterministic models as they solve the Navier-Stokes equation and can be used to explicitly calculate turbulences at very fine grid resolution and in complex geometries, although they require large computational power (Leelőssy et al. 2014)

1.2 Problem Statement and Context

The potential uses of deterministic air quality models are all very important and convenient, for example, in the assessment of emission change impacts, identification of source impacts, support of potential regulatory direction changes, evaluation of impacts of climate change on surface level air quality, design of monitoring and observation networks, and providing spatial pollutant fields for health and exposure assessments. However, the uncertainty inherent in emissions

inventories has significantly reduced the models´ ability to diagnose and forecast air quality and its impacts (Lindley et al., 2000). Thus, of the entire atmospheric modeling system, the generation of emission inventories remains the process that (by far) provides the greatest uncertainty in the applications of such modeling systems.

This is relevant since the emission control strategies that arise from the use and analysis of the results derived from air quality models could be erroneous by employing an emissions inventory that is intrinsically flawed, or the associated uncertainties are so considerable that it is not possible to decide based on the results of the modeling.

Emission inventories are usually developed from the bottom-up method, which is based on the analysis of historical statistics by activity (such as energy consumption and industrial production) or on emission factors specific to source or region. Although inventories constructed in this way represent the direct method for estimating existing emissions, it is difficult to calculate the characteristic uncertainty of statistics, emission factors, random errors in measurement devices, poorly determined processes, time profiles, and spatial placement factors. Moreover, estimating current emissions by this method is particularly difficult because the publication of the statistics necessary for their calculation usually presents a lag period of between two and five years with respect to their sampling and/or calculation (Cheng et al. 2021). For example, in Mexico, national criteria pollutant emission inventories have been published for base years 2005, 2008, 2013 and 2016, only (Secretaría de Medio Ambiente y Recursos Naturales, 2022).

1.4 Solution Overview

One way to reduce the uncertainty of the output results of air quality models is to use inverse modeling or data assimilation to improve the quality of information provided in emissions inventories. The general idea of this technique is to use data measured of ambient air pollutants to feed back into the air quality model and "force" it to properly reproduce the concentration fields by modifying certain input

data or parameters of the model, typically those that are considered the most uncertain. See Figure 1.3 for details.

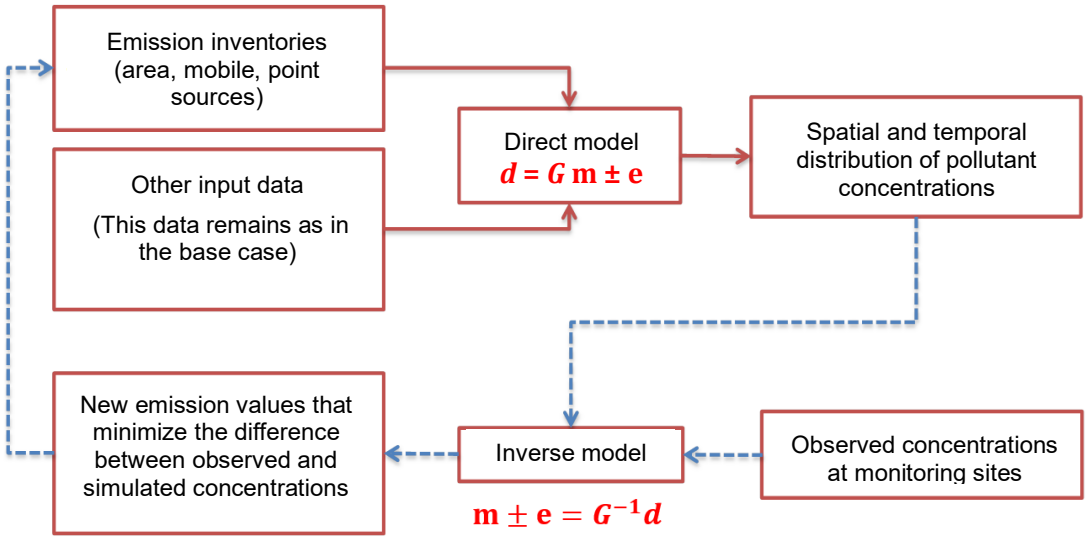

Figure 1.3. Inverse modeling approach for a photochemical modeling application.

There are numerous applications of using inverse modeling to suggest changes in emission inventories, from global (Pétron et al., 2002) or continental (Elbern et al., 2007) scales to local or urban scales (Quélo et al., 2005). Examples of these applications include: estimation of NO_x emissions from satellite observations (Jaeglé et al., 2005; Martin et al., 2006; Wang et al., 2007), the assimilation of chemical concentrations in the calculation of inventories of O_3 and reactive nitrogen (Pierce et al., 2007), the estimation of NH_3 emissions in the United States, and the modeling of dynamic particle growth by condensation (Henze et al., 2004; Sandu et al., 2005).

In this context, assimilation techniques such as three-dimensional variational methods (3DVAR) (Barker et al., 2004; Li et al. 2013; Hu et al. 2022) or four dimensions (4DVAR) (e.g. Meirink et al., 2008; Cao et al. 2020), the Kalman filter (e.g. Kong et al. 2019), or the use of the air quality model adjunct (e.g. Hakami et al., 2005) have been applied in solving inverse modeling problems. Additionally, techniques such as Green's function (Dougherty and Rabbitz 1979; Kramer et al., 1984), Automatic Differentiation in Fortran (ADIFOR) (Bischof et al., 1992), and the

Direct Decoupled Method (DDM) (Dunker, 1984; Dunker et al. 2002, Arter and Arunachalam, 2021) have been used in the sensitivity analysis of these same air quality models. Recently, Cheng et al. (2021) developed a new inversion method for emission sources modeling, based on the three-dimensional decoupled direct (DDM-3D) sensitivity analysis module in the Community Multiscale Air Quality (CMAQ) model and the 3DVAR data assimilation technique.

Although most applications use the variational approach (3DVAR or 4DVAR), it is necessary to make some modifications to traditional methods, such as solving the minimization function with constraints, in such a way as to ensure that we obtain positive solutions (there can be no negative emissions), balance errors or add terms that taken account for co-location of sources and observations (Saide et al., 2009). Another approach is performing inverse modeling from regularization techniques, a formal method that has been little explored.

Given the above, this work aims to explore the use of mathematical tools of inverse modeling based on matrix regularization in conjunction with air quality models with the aim of studying their possible advantages in applications that seek to identify possible improvements in the emission inventories required by the air quality model, and thus reduce the uncertainty in these emissions and improve the overall performance of the atmospheric modeling system.

On the other hand, with the increasing availability of large amounts of historical data, it has been possible to complement the benefits of deterministic models with the use of empirical models to answer questions related to air pollution. MLR models are globally and widely used over many years as a method for air pollution forecasting, which can help to attempt the uncertainty of the future simply by relying on past and current data for decision-making.

1.3 Research Questions

In this work, the following research questions are addressed:

1. Does uncertainty exist in emission inventories in Mexican metropolitan areas?
2. Are there significant differences among emission inventories in Mexican Metropolitan Areas?
3. Can a deterministic quality model accurately reproduce pollutant concentrations in a Mexican metropolitan area?
4. Are matrix regularization techniques adequate tools for improving air emissions inventories?
5. Does the use of mathematically “corrected” emission inventories improve air quality models performance?
6. What energy and economic indicators are structurally related to air pollution in Mexican cities?
7. Can multivariate linear regression models based on energy and economic variables accurately represent air quality in Mexican metropolitan areas?
8. What are the main economic activities influencing air quality in Mexican metropolitan areas?

1.4 Objectives

The purpose of this research is to explore data-driven based tools, in specific inverse modeling mathematical tools to improve emission inventories for air quality modeling applications, and to quantify the contribution of energy and economic factors to air pollution.

The specific objectives are:

1. To identify inverse modeling mathematical tools based on matrix regularization that can be implemented in conjunction with air quality models to improve emission inventories and air quality models performance.
2. To build and assess multivariate linear regression models based on energy and economic variables to estimate air quality in Mexican metropolitan areas.

1.6 Dissertation Organization

This dissertation is organized as follows:

This **Chapter 1** serves as an introduction to the problem object of this thesis. It describes what an air quality model consists of, its advantages and disadvantages, and the information requirements they have, emphasizing the importance of having detailed and accurate emissions inventories. It also introduces the possibility of using inverse modeling and matrix regularization tools for the mathematical correction of currently available inventories, and reflects on the use of probabilistic models, such as Multivariate Linear Regression Models, as complementary tools to deterministic models, to better understand urban air pollution main drivers and emission sources.

Chapter 2 presents an exploratory analysis of existing air pollution inventories for criteria pollutants in Mexico, specifically for the Monterrey Metropolitan Area, and discusses on the differences among inventories produced by different agencies, and their implications for air quality modeling applications. A study case is described, in which the statistical performance of the Community Multiscale Air Quality Model (CMAQ) is tested using different emission inventories for a modeling domain centered in the Monterrey Metropolitan Area.

Chapter 3 consists of a review of matrix regularization methods and their applications in atmospheric sciences. The chapter presents an overview of existing regularization methodologies, with special emphasis on direct methods, such as Tikhonov regularization, ridge regression, truncated singular value decomposition, and damped singular decomposition, which differ from each other according to the mathematical form that this functional has. It also briefly discusses some of the most common regularization parameter selection methods, such as the discrepancy principle, the unbiased predictive risk estimator, generalized cross-validation, the L-curve method, normalized cumulative periodograms and hybrid methods, which are briefly described. It also presents the concept of regularization with restrictions and describes some efforts in this regard. Finally, as an example, some applications of direct regularization methods in the field of atmospheric sciences are listed.

Chapter 4 is a preliminary application of inverse modeling and some of the methods identified in Chapter 3, in conjunction with an exploratory air quality model, to identify that combination of regularization method and regularization parameter selection technique that identifies and corrects more rightly the possible errors within the emissions inventory. This preliminary application is made for a domain centered in the Guadalajara Metropolitan Area.

Chapter 5 investigates the regional air quality characteristics and its drivers in three Mexican metropolitan areas: Monterrey, Guadalajara and Mexico City, and analyzes the spatial and temporal characteristics of air quality, as well as the influence of energy and economic variables, routinely recorded by government agencies on a monthly basis, by using statistical analysis methods. The study provides a theoretical basis for identifying and quantifying the causes of urban air pollution and allows the formulation of pollution control measures in Mexico. (Yang et al., 2022).

Finally, **Chapter 6** presents the conclusions and recommendations of this study.

1.7 References

Abdullah, S., Napi, N. N. L. M., Ahmed, A. N., Mansor, W. N. W., Mansor, A. A., Ismail, M., ... & Ramly, Z. T. A. (2020). Development of multiple linear regression for particulate matter (PM10) forecasting during episodic transboundary haze event in Malaysia. *Atmosphere*, *11*(3), 289.

Arter, C. A., & Arunachalam, S. (2021). Assessing the importance of nonlinearity for aircraft emissions' impact on O3 and PM2. 5. *Science of the Total Environment*, *777*, 146121.

Bai, L., Wang, J., Ma, X., & Lu, H. (2018). Air pollution forecasts: An overview. *International journal of environmental research and public health*, *15*(4), 780.

Barker, D. M., Huang, W., Guo, Y. R., Bourgeois, A. J., & Xiao, Q. N. (2004). A three-dimensional variational data assimilation system for MM5:

Implementation and initial results. *Monthly Weather Review*, *132*(4), 897-914.

Bischof, C., Carle, A., Corliss, G., Griewank, A., & Hovland, P. (1992). ADIFOR–generating derivative codes from Fortran programs. *Scientific Programming*, *1*(1), 11-29.

Cao, H., Henze, D. K., Shephard, M. W., Dammers, E., Cady-Pereira, K., Alvarado, M., ... & Edgerton, E. S. (2020). Inverse modeling of NH3 sources using CrIS remote sensing measurements. *Environmental Research Letters*, *15*(10), 104082.

Carmona, J. M., Gupta, P., Lozano-García, D. F., Vanoye, A. Y., Hernández-Paniagua, I. Y., & Mendoza, A. (2021). Evaluation of MODIS aerosol optical depth and surface data using an ensemble modeling approach to assess PM2. 5 temporal and spatial distributions. *Remote Sensing*, *13*(16), 3102.

Cheng, X., Hao, Z., Zang, Z., Liu, Z., Xu, X., Wang, S., ... & Ma, X. (2021). A new inverse modeling approach for emission sources based on the DDM-3D and 3DVAR techniques: an application to air quality forecasts in the Beijing–Tianjin–Hebei region. *Atmospheric Chemistry and Physics*, *21*(18), 13747-13761.

Dougherty, E. P., & Rabitz, H. (1979). A computational algorithm for the Green's function method of sensitivity analysis in chemical kinetics. *International journal of chemical kinetics*, *11*(12), 1237-1248.

Dunker, A. M. (1984). The decoupled direct method for calculating sensitivity coefficients in chemical kinetics. *The Journal of chemical physics*, *81*(5), 2385-2393.

Dunker, A. M., Yarwood, G., Ortmann, J. P., & Wilson, G. M. (2002). The decoupled direct method for sensitivity analysis in a three-dimensional air quality model implementation, accuracy, and efficiency. *Environmental Science & Technology*, *36*(13), 2965-2976.

Elbern, H., Strunk, A., Schmidt, H., & Talagrand, O. (2007). Emission rate and chemical state estimation by 4-dimensional variational inversion. *Atmospheric Chemistry and Physics*, *7*(14), 3749-3769.

inversion," Atmospheric Chemistry and Physics, vol. 7, pp. 3749-3769, 2007.

Environmental Protection Agency. (2022, July 11). Managing Air Quality - Air Quality Modeling. Retrieved November 8, 2022, from https://www.epa.gov/air-quality-management-process/managing-air-quality-air-quality-modeling

Ganesh, S. S., Modali, S. H., Palreddy, S. R., & Arulmozhivarman, P. (2017, May). Forecasting air quality index using regression models: A case study on Delhi and Houston. In *2017 International Conference on Trends in Electronics and Informatics (ICEI)* (pp. 248-254). IEEE.

Ganesh, S. S., Arulmozhivarman, P., & Tatavarti, R. (2019). Forecasting air quality index using an ensemble of artificial neural networks and regression models. *Journal of Intelligent Systems*, *28*(5), 893-903.

Grell, G. A., Dudhia, J., & Stauffer, D. R. (1994). A description of the fifth-generation Penn State/NCAR Mesoscale Model (MM5).

Hakami, A., Henze, D. K., Seinfeld, J. H., Chai, T., Tang, Y., Carmichael, G. R., & Sandu, A. (2005). Adjoint inverse modeling of black carbon during the Asian Pacific Regional Aerosol Characterization Experiment. *Journal of Geophysical Research: Atmospheres*, *110*(D14).

He, Z., Liu, P., Zhao, X., He, X., Liu, J., & Mu, Y. (2022). Responses of surface O3 and PM2. 5 trends to changes of anthropogenic emissions in summer over Beijing during 2014–2019: A study based on multiple linear regression and WRF-Chem. *Science of The Total Environment*, *807*, 150792.

Henze, D. K., Seinfeld, J. H., Liao, W., Sandu, A., & Carmichael, G. R. (2004). Inverse modeling of aerosol dynamics: Condensational growth. *Journal of Geophysical Research: Atmospheres*, *109*(D14).

Jaeglé, L., Steinberger, L., Martin, R. V., & Chance, K. (2005). Global partitioning of NO x sources using satellite observations: Relative roles of fossil fuel combustion, biomass burning and soil emissions. *Faraday discussions*, *130*, 407-423.

Juda-Rezler, K. (1991). Classification and characteristics of air pollution models. In *Chemistry for the Protection of the Environment* (pp. 51-72). Springer, Boston, MA.

Kong, L., Tang, X., Zhu, J., Wang, Z., Pan, Y., Wu, H., ... & Carmichael, G. (2019). Improved inversion of monthly ammonia emissions in China based on the Chinese ammonia monitoring network and ensemble Kalman filter. *Environmental Science & Technology*, *53*(21), 12529-12538.

Kramer, M. A., Rabitz, H., Calo, J. M., & Kee, R. J. (1984). Sensitivity analysis in chemical kinetics: Recent developments and computational comparisons. *International Journal of Chemical Kinetics*, *16*(5), 559-578.

Leelőssy, Á., Molnár, F., Izsák, F., Havasi, Á., Lagzi, I., & Mészáros, R. (2014). Dispersion modeling of air pollutants in the atmosphere: a review. Open Geosciences, 6(3), 257-278.

Li, Z., Zang, Z., Li, Q. B., Chao, Y., Chen, D., Ye, Z., ... & Liou, K. N. (2013). A three-dimensional variational data assimilation system for multiple aerosol species with WRF/Chem and an application to PM 2.5 prediction. *Atmospheric Chemistry and Physics*, *13*(8), 4265-4278.

Li, X., Hussain, S. A., Sobri, S., & Said, M. S. M. (2021). Overviewing the air quality models on air pollution in Sichuan Basin, China. *Chemosphere*, 271, 129502.

Liao, K., Huang, X., Dang, H., Ren, Y., Zuo, S., & Duan, C. (2021). Statistical approaches for forecasting primary air pollutants: a review. *Atmosphere*, *12*(6), 686.

Maçaira, P. M., Thomé, A. M. T., Oliveira, F. L. C., & Ferrer, A. L. C. (2018). Time series analysis with explanatory variables: A systematic literature review. *Environmental Modelling & Software*, *107*, 199-209.

Martin, R. V., Sioris, C. E., Chance, K., Ryerson, T. B., Bertram, T. H., Wooldridge, P. J., ... & Flocke, F. M. (2006). Evaluation of space-based constraints on global nitrogen oxide emissions with regional aircraft measurements over and downwind of eastern North America. *Journal of Geophysical Research: Atmospheres*, *111*(D15).

Meirink, J. F., Bergamaschi, P., & Krol, M. C. (2008). Four-dimensional variational data assimilation for inverse modelling of atmospheric methane emissions: method and comparison with synthesis inversion. *Atmospheric chemistry and physics*, *8*(21), 6341-6353.

Napelenok, S. L., Foley, K. M., Kang, D., Mathur, R., Pierce, T., & Rao, S. T. (2011). Dynamic evaluation of regional air quality model's response to emission reductions in the presence of uncertain emission inventories. *Atmospheric Environment*, *45*(24), 4091-4098.

Osisanwo, F. Y., Akinsola, J. E. T., Awodele, O., Hinmikaiye, J. O., Olakanmi, O., & Akinjobi, J. (2017). Supervised machine learning algorithms: classification and comparison. *International Journal of Computer Trends and Technology (IJCTT)*, *48*(3), 128-138.

Pantusheva, M., Mitkov, R., Hristov, P. O., & Petrova-Antonova, D. (2022). Air Pollution Dispersion Modelling in Urban Environment Using CFD: A Systematic Review. *Atmosphere*, *13*(10), 1640.

Pétron, G., Granier, C., Khattatov, B., Lamarque, J. F., Yudin, V., Müller, J. F., & Gille, J. (2002). Inverse modeling of carbon monoxide surface emissions using Climate Monitoring and Diagnostics Laboratory network observations. *Journal of Geophysical Research: Atmospheres*, *107*(D24), ACH-10.

Pierce, R. B., Schaack, T., Al-Saadi, J. A., Fairlie, T. D., Kittaka, C., Lingenfelser, G., ... & Fishman, J. (2007). Chemical data assimilation estimates of continental US ozone and nitrogen budgets during the Intercontinental Chemical Transport Experiment–North America. *Journal of Geophysical Research: Atmospheres*, *112*(D12).

Quélo, D., Mallet, V., & Sportisse, B. (2005). Inverse modeling of NOx emissions at regional scale over northern France: Preliminary investigation of the second-order sensitivity. *Journal of Geophysical Research: Atmospheres*, *110*(D24).

Rosenlund, M., Forastiere, F., Stafoggia, M., Porta, D., Perucci, M., Ranzi, A., ... & Perucci, C. A. (2008). Comparison of regression models with land-use and

emissions data to predict the spatial distribution of traffic-related air pollution in Rome. *Journal of exposure science & environmental epidemiology*, *18*(2), 192-199.

Saide, P., Osses, A., Gallardo, L., & Osses, M. (2009). Adjoint inverse modeling of a CO emission inventory at the city scale: Santiago de Chile's case. *Atmospheric Chemistry and Physics Discussions*, *9*(2), 6325-6361.

Sandu, A., Daescu, D. N., Carmichael, G. R., & Chai, T. (2005). Adjoint sensitivity analysis of regional air quality models. *Journal of Computational Physics*, *204*(1), 222-252.

Secretaría de Medio Ambiente y Recursos Naturales. (2019, March 22). Inventario Nacional de Emisiones de Contaminantes Criterio INEM. Retrieved November 8, 2022, from https://www.gob.mx/semarnat/acciones-y-programas/inventario-nacional-de-emisiones-de-contaminantes-criterio-inem#:~:text=El%20Inventario%20Nacional%20de%20Emisiones,SOx)%20y%20part%C3%ADculas%20con%20di%C3%A1metro

Shams, S. R., Jahani, A., Kalantary, S., Moeinaddini, M., & Khorasani, N. (2021). The evaluation on artificial neural networks (ANN) and multiple linear regressions (MLR) models for predicting SO2 concentration. *Urban Climate*, *37*, 100837.

Skamarock, W. C., Klemp, J. B., Dudhia, J., Gill, D. O., Barker, D. M., Wang, W., & Powers, J. G. (2005). *A description of the advanced research WRF version 2*. National Center For Atmospheric Research Boulder Co Mesoscale and Microscale Meteorology Div.

Taheri Shahraiyni, H., & Sodoudi, S. (2016). Statistical modeling approaches for PM10 prediction in urban areas; A review of 21st-century studies. *Atmosphere*, *7*(2), 15.

Wang, Y., McElroy, M. B., Martin, R. V., Streets, D. G., Zhang, Q., & Fu, T. M. (2007). Seasonal variability of NOx emissions over east China constrained by satellite observations: Implications for combustion and microbial sources. *Journal of Geophysical Research: Atmospheres*, *112*(D6).

Wilks, D. S. (2011). Statistical methods in the atmospheric sciences (Vol. 100). Academic press.

Yarwood, G., Rao, S., Yocke, M. A., & Whitten, G. (2005). Updates to the carbon bond chemical mechanism: CB05. *Final report to the US EPA, RT-0400675*, 8, 13.

Yu, S., Mathur, R., Sarwar, G., Kang, D., Tong, D., Pouliot, G., & Pleim, J. (2010). Eta-CMAQ air quality forecasts for O 3 and related species using three different photochemical mechanisms (CB4, CB05, SAPRC-99): comparisons with measurements during the 2004 ICARTT study. *Atmospheric Chemistry and Physics*, *10*(6), 3001-3025.

Chapter 2

Analysis of anthropogenic emission inventories available for the Monterrey Metropolitan Area

2.1 Introduction

An emission inventory is a structured set of emission data distinguishing different pollutants and source categories, for a certain geographical location and period, and is a key input to deterministic air quality models, as it provides the chemical forcing component (Pouliot et al. 2015). However, emission inventories have been identified as one of the most important sources of uncertainty in air quality modeling applications (Guevara et al. 2017).

The estimation of emissions requires information concerning activity factors (e.g. total amount of fuel consumed) and emission factors per activity (e.g. amount of pollutant emitted per activity unit), and can be computed as the product of emission factors times the activity data:

$$\boldsymbol{EI} = \sum_{\boldsymbol{i}} (\boldsymbol{EF_i} \times \boldsymbol{AF_i}) \qquad \textbf{(2.1)}$$

Where $\boldsymbol{EI}$ is the total emission inventory, $\boldsymbol{EF_i}$ is the emission factor for the emissions of a given pollutant from source category i, and $\boldsymbol{AF_i}$ is the activity factor for source category i. Emission factors and activity data are classified into three tiers to differentiate their reliability and methodological complexity. Tier 1 is the basic method, while Tier 3 involves the use of the most specific data to produce more accurate emission estimates.

There are two main approaches for estimating emission inventories: bottom-up and top-down methods, both of which have have advantages and disadvantages. For

example, bottom-up approaches rely on specific information for each sector or source category, allowing for higher spatial and temporal detail but requiring larger amounts of data and resources. For example, local emission inventories are usually constructed using bottom-up methodologies, based on local energy and fuel consumption data which are aggregated to the required spatial scale (e.g., Hestia et al. 2019).

On the other hand, top-down approaches are based on the disaggregation of variables defined at regional or national level in smaller areas based on variables thar serve as proxies for specific activities. National Inventory Reports (NIR) (UNFCCC, 2019) are country-level, yearly emissions typically estimated through a top-down approach, based on energy statistics. NIR can be spatially and temporally disaggregated (scaled-down) to a certain level using proxies. Ultimately, the selection of method depends on the emission source considered, data availability and spatial coverage (Guevara et al. 2017).

As described above, the generation of emission inventories is a complex task that requires capacity building, collection of large amounts of data and development of methodologies for the estimation of emissions from different sources. Therefore, inventories are not updated frequently, and commonly, the methodologies used to build them are revised and changed from one period to another. For these reasons, during the compilation of an emission inventory, uncertainties are introduced at different levels (e.g., methodology, magnitude, timing, locations), and increasingly more attention is given to this topic.

For greenhouse gases inventories, parties of the United Nations Framework Convention on Climate Change (UNFCCC) require that NIR include an assessment of the uncertainties in the underlying data and an analysis of the uncertainties in the total emissions following Intergovernmental Panel on Climate Change (IPCC) guidelines. The simplest uncertainty analysis is based on simple equations for combining uncertainties from different sources, e.g., Tier 1 approach.

A more advanced approach is a Monte Carlo simulation, which allows for non-normal uncertainty distributions (Tier 2 approach). Monte Carlo simulations for uncertainty analysis take as inputs the uncertainty distribution for each variable and an equation for the calculation of a desired quantity. The desired quantity is repeatedly (usually, more than 100 times) calculated by randomly drawing from the specified uncertainty distributions of the input variables, with new random drawings each time, until a resulting uncertainty distribution of the calculated value is obtained (Albert, 2020). The Tier 2 approach has been used for estimating uncertainty of national emission inventories of Finland (Monni et al., 2004) and Denmark (Fauser et al., 2011, Super et al. 2020).

The uncertainties in emission inventories are important to understand for several reasons. First, knowledge of uncertainties helps to pinpoint emission sources or areas that require more scrutiny (Monni et al., 2004). Second, knowledge of uncertainties in prior emission estimates is an important part of inverse modelling frameworks, which can be used for emission verification and in support of decision-making (Andres et al., 2014). If uncertainties are not properly considered, there is a risk that the uncertainty range does not contain the actual emission value. In contrast, if uncertainties are overestimated, the initial emission inventory gives little information about the actual emissions and more independent observations are needed.

Although the estimation of uncertainty intervals remains a complex task, some emission inventories have tried to estimate them. For example, Streets et al. (2003) estimated that the uncertainties of Chinese emissions varied from −12% ~13% for SO2 to −83% ~495% for organic carbon (expressed as the lower and upper bounds of a 95% confidence interval, CI, around a central estimate). Zhao et al. (2011) reported uncertainties (95% CI around the central estimates) of Chinese emissions of SO_2, NO_x, total PM, PM_{10}, $PM_{2.5}$, black carbon (BC), and organic

carbon (OC) in 2005 to be −14%~13%, −13%~37%, −11%~38%, −14%~45%, −17%~54%, −25%~136%, and −40%~121%, respectively.

In Mexican inventories, emission estimates for specific sources have also reported large degrees of uncertainty. For example, when preparing the BRAVO emissions inventory (BRAVO-EI), Southern California Edison and the US Environmental Protection Agency estimated that the Carbón power plant -the largest SO_2 source in the state of Coahuila, Mexico, emitted approximately 241,000 ton/year of SO_2, which differed in 58.5% from the 152,000 ton/year estimate provided by Secretariat of Energy of Mexico (Secretaría de Energía) (Kuhns et al. 2005). Discrepancies also arose once the National Emission Inventory for the Six Norther Border States of Mexico (NBS-EI) was completed in April 2004 and compared to the 1999 BRAVO-EI. Overall, SO_2 emissions were a factor of 1.4 lower in the BRAVO-EI than in the NBS-EI, for regions outside of Coahuila. Emissions of NH_3 and PM_{10} were lower in the BRAVO EI than in the NBS-EI by 15 and 32%, respectively. However, with regards to CO and VOC emissions, the BRAVO-EI values exceeded the NBS-EI values by factors of 3.8 and 1.8, respectively, and NO_x and $PM_{2.5}$ emissions were within 10% of each other for the two inventories (Kuhn et al. 2005). The use of different estimation methodologies and source classification criteria (e.g. point and area sources) may explain some of the exhibited differences (Kuhn et al. 2005).

The BRAVO-EI was also compared with non-U.S. emissions from the EDGAR v3 Emission Database for Global Atmospheric Research (EDGARv3-EI), an anthropogenic emissions inventory generated for global atmospheric and global climate modeling. Details can be found in Kuhn et al. (2005). The exercise focused on SO_2 emissions, and concluded that the EDGARv3-EI SO_2 emissions in the 10 northernmost Mexico states within the BRAVO-EI were 70% higher than the BRAVO-EI.

Regarding greenhouse gases (GHG) emissions estimations, most of Mexico's emissions have been calculated using the default Tier 1 methods from the Intergovernmental Panel on Climate Change (IPCC) with generic emission factors, which entails a large degree of uncertainty for some sectors. For example, Mexico's methane emissions from oil and gas have 56% and 40% uncertainty, respectively, according to the national inventory (Scarpelli et al. 2020). Overall, the Instituto Nacional de Ecología y Cambio Climático reports uncertainty estimates for national emissions by subsector as ±2σ relative error standard deviations (INECC, 2018).

More recently, Scarpelli et al. (2020) prepared a gridded inventory of Mexico's anthropogenic methane emissions for 2015 with 0.1°×0.1° resolution (≈10×10 km^2) and detailed sectoral breakdown. One of the major findings was that large differences between their inventory and previous gridded emission inventories for Mexico, in particular EDGAR v5, existed. EDGAR v5 estimated total CH_4 emissions 25% higher than the national inventory. Particularly, solid waste disposal emissions were a factor of two greater, coal emissions 81% lower, and oil/gas emissions 78% higher. Of interest, the Mexico City GHG 2016 Emission Inventory reported that 70% of CH_4 emissions originated from solid waste, consistent with Scarpelli et al.´s CH_4emission inventory (74%), while EDGAR v5 reported 74% of its emission from wastewater.

2.2 Research framework

This section presents a comparison of the emission inventories published for the Metropolitan Area of Monterrey (MMA) and the state of Nuevo León with the aim of discerning if significant differences exist among Nuevo León´s emissions inventories and their suitability for air quality modeling applications. This activity was carried out in three stages: A first stage, where the inventories available until the base year 2005 were explored and compared one to another.

In a second stage, as study case, a modeling application centered in the Monterrey Metropolitan Area, was executed to analyze the implications of using different emission inventories on an air quality model´s response. The results of stages 1 and 2 served as justification for the development of the following sections of this dissertation, including inverse modeling applications.

However, a third stage, namely an update of the inventory analysis was later carried out to consider inventories developed for base years after 2005, and thus confirm the ongoing relevance of this research. See Figure 2.1.

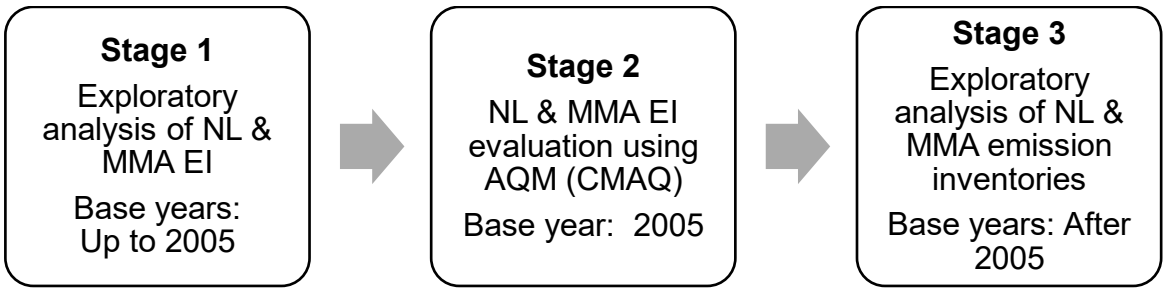

Figure 2.1. Research framework for Chapter 2 of this dissertation.

2.3 Area and period of study

The state of Nuevo Leon is home to 5,784,442 people, of which 5,341,171 (92.3%) live in the Monterrey Metropolitan Area (MMA) (INEGI, 2020), encompassing 18 municipalities, namely, Abasolo, Apodaca, Cadereyta Jiménez, El Carmen, Ciénega de Flores, García, San Pedro Garza García, General Escobedo, General Zuazua, Guadalupe, Juárez, Monterrey, Pesquería, Salinas Victoria, San Nicolás de los Garza, Hidalgo, Santa Catarina and Santiago.

Despite being considered the industrial capital of Mexico, the second most populous metropolitan area in the country, and listed as the second-most polluted city in Latin America (e.g. Gouveia et al. 2021), there is a lack of air quality modeling studies for the MMA. Some isolated modeling efforts are Vivanco-Moreno

and Ramírez-Lara (2007), who applied the HYSPLIT model to evaluate air trajectories and their impact on the dispersion of pollutants to atmospheric pollutants; and Mendoza-Domínguez (1996, 2000), who utilized the California Institute of Technology (CIT) (Russell et al. 1988, Harley et al. 1993) photochemical model to study the impact of anthropogenic and biogenic emissions in the Mexico-United States border area.

Later, Sierra et al. (2013) applied the Community Multiscale Air Quality (CMAQ) photochemical model to simulate environmental chemistry and transport of pollutants – particularly ozone – during an episode of high pollution in northeastern Mexico, incorporating a broad region that encompassed sections of the states of Nuevo León, Coahuila and Tamaulipas, as well as southern Texas in the United States. However, although the model favorably simulated ozone concentrations within the domain, particular situations were identified where model performance was unfavorable, presumably due to the influence of the topography and box resolution used (8×8 km) (Sierra et al. 2013).

2.4 Emission Inventory

2.4.1 Comparison of emission inventories, up to base year 2005.

Table 2.1 presents the criteria pollutants inventories analyzed in Stage 1 of this study. The revised inventories were: National Inventory of Emissions of Mexico 1999 (SEMARNAT, 2006), the Program to Improve the Air Quality of the Metropolitan Area of Monterrey PROAIRE 2008-2012 (Gobierno del Estado de Nuevo León, 2009), and the National Atmospheric Emissions System (SINEA, 2011), as well as the results condensed in the document "Exploitation of the results of the National Inventory of Emissions of Mexico INEM 2005" (Landa Fonseca, 2012). These inventories were publicly available with base years up to 2005 and had a spatial coverage of the MMA and/or the State of Nuevo León.

Table 2.1. Emission inventories considered in this study.

Acronym	Emission Inventory	Author	Base Year	Year of publication
INEM99	Mexico National Emissions Inventory 1999	SEMARNAT-INE	1999	2006
PROAIRE05	Management Program to Improve Air Quality of the Metropolitan Area of Monterrey 2008-2012	Gobierno del Estado de Nuevo León-SEMARNAT	2005	2009
SINEA05	National System of Emissions to the Atmosphere	SEMARNAT	2005	2011
INEM05	National Emissions Inventory of Mexico 2005	SEMARNAT-INE	2005	2011-2012
LANDA05	Exploitation of the results of the National Inventory of Emissions of Mexico INEM 2005	Landa Fonseca	2005	2012

Emission inventories consider different categories of sources: mobile, stationary, area, and natural. A direct comparison was made among the emission values reported per category for the inventories listed in Table 2.1. Because the national inventories prepared by SEMARNAT were built with a “bottom-up” approach, in which state emissions were either estimated through aggregating municipal data or disaggregating state data into municipal resolution, two spatial domains were considered: State of Nuevo Leon, which would include emissions generated in the whole state, and the MMA, which only included emissions from the municipalities geographically located within the MMA. Tables 2.2 and 2.3 show the comparison for mobile and area sources.

For example, for the AMM, SINEA05 reported gaseous emissions from mobile sources, up to almost 6 times higher than those previously reported in PROAIRE05 (Table 2.2). It is worth mentioning that in the case of mobile sources, the results published in SINEA05 and those reported in LANDA05 were equivalent. The different inventories prepared for area sources within the MMA also present severe discrepancies with each other (Table 2.3). For area sources, the most substantial difference corresponds to the estimates of SO_2 emissions of SINEA05 with respect

to those published in PROAIRE05 where the former are almost 800 times greater than the latter, going from 23.3 to 18,187.9 tons of SO_2. This large difference can be attributed to the emissions of SO_2 reported from industrial combustion of fuel oil in Monterrey (4995 ton), Apodaca (3148 ton), San Nicolás (2885 ton), Guadalupe (2082 ton), Santa Catarina (1879 ton), in addition to the smaller contributions of other municipalities. On the other hand, although both INE1999 and PROAIRE contain additional categories of point sources and mobile non-road mobile, SINEA does not explicitly include them.

Table 2.2 Comparison of emission inventories for on-road mobile sources (ton/year).

Coverage	MMA			State of Nuevo León		
Inventory	SINEA05	PROAIRE05	INEM99	INEM99	SINEA05	LANDA05
NO_X	164,410	31,762	39,209	40,350	177,206	177,206
SO_2	1,694	879	2,080	2,145	1,831	1,831
COV	248,126	51,868	51,081	52,458	282,848	282,848
CO	3,315,153	491,863	380,366	391,398	3,783,994	3,783,994
PM_{10}	903	818	1,749	1,804	969	969
$PM_{2.5}$	546	568	1,603	1,653	584	584
NH_3	1,565	ND	605	621	1,696	1,696

Table 2.3 Comparison of emission inventories for area sources (ton/year).

Coverage	MMA			State of Nuevo León		
Inventory	SINEA05	PROAIRE05	INEM99	INEM99	SINEA05	LANDA05
NO_X	7,372	2,364	5,847	7,699	8,167	8,840
SO_2	18,188	23	16,068	17,357	19,576	19,688
COV	65,708	47,752	61,294	72,793	75,266	73,723
CO	25,682	2,499	14,825	25,929	41,560	105,601
PM_{10}	8,360	42,790	2,853	5,334	13,532	24,021
$PM_{2.5}$	4,757	9,414	1,992	3,558	7,473	16,025
NH_3	19,097	ND	4,800	24,847	28,091	776

2.5 Air quality modeling

Because AMM emission estimates presented large differences among the available inventories, an air quality modeling experiment was instrumented to *i*) evaluate the performance of an air quality model under different emission scenarios, and *ii*) identify the "best" emission inventory for air quality modeling applications in the AMM among a set of alternative emission inventories.

The Models-3 Computational Framework, integrated by the Fifth Generation Meteorological Model (MM5) (Grell et al. 1994), Sparse Matrix Operator Kernel System (SMOKE) (Houyoux and Vukovich, 1999), and Community Multiscale Air Quality (CMAQ) chemical transport model was chosen as the modeling system, and will be described in the following sections (Figure 2.2).

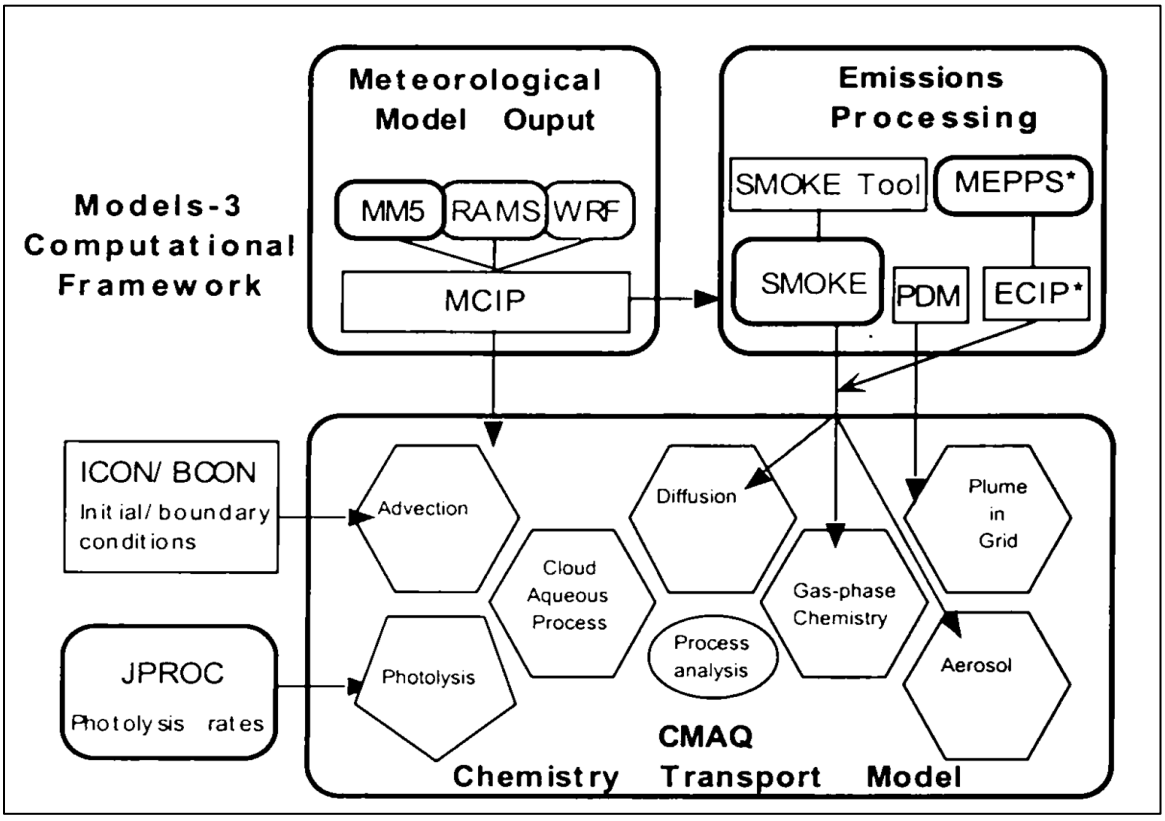

Figure 2.2 Modules required to run the CMAQ Models-3 Computational Framework (From: Byun and Shere, 2006)

2.5.1. Emission inventory processing

Three-dimensional, photochemical, transport models, such as CMAQ require detailed, spatially-gridded emission data. The SMOKE (Houyoux and Vukovich, 1999), emission processor engine was developed in the United States and redesigned by the US Environmental Protection Agency to support air quality modelling activities. Emission data is then chemically speciated, temporally allocated, spatially gridded, and merged into model-ready emissions (see Figure 2.3). Further details on the SMOKE processor can be found in CMAS (2022).

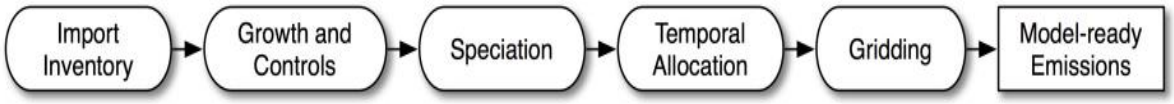

Figure 2.3 SMOKE methodology. (From CMAS, 2022).

At the timing of this research, INE99 was the only AQM-formatted emission inventory, suitable for SMOKE processing. Therefore, the INE99 emission inventory was selected as the base inventory to be updated to year 2005 with the help of the CNTLMAT tool, which generates a matrix of growth factors for all emission sources, from which revised inventories are generated.

Thus, based in the analysis of section 2.4.1, the scaling factors that would be applied to the INE99 emission inventory to update it to the year 2005 were calculated. These scaling factors are, in other words, the percentages in which the emissions reported in the INE99 needed to be increased or decreased to total the emissions reported in the different inventories. Table 2.4 presents a summary of the scaling factors that were defined for the different inventories (or combinations thereof) that were evaluated. As can be seen, in most cases NH_3 emissions were not escalated due to insufficient data. The case INE99-SINEA05/PROAIRE05/LANDA05 consists of using SINEA05 for area and mobile sources that circulate on the road, PROAIRE05 for non-road mobile sources and the emissions reported in LANDA05 for stationary sources.

It is important to remember that the use of these scaling factors in updating emissions within the modeling domain is subject to the "gross assumption that the entire Northeast region had socioeconomic dynamics similar to that experienced by the AMM in the period 1999 to 2005" (Sierra, 2011).

Table 2.4 Scale factors calculated for the different inventories available.

Scale-up factors for emission inventories INE99-PROAIRE05							
Source	NO_X	SO_2	VOC	CO	PM_{10}	$PM_{2.5}$	NH_3
Area	-0.59577	-0.99856	-0.22095	-0.83147	13.99876	3.72643	0.00000
Mobile	-0.18993	-0.57735	0.01540	0.29313	-0.53232	-0.64557	0.00000
Non-road mobiles	-0.05984	-0.06241	0.02695	-0.03828	-0.02860	-0.02901	0.00000
Fixed	0.01785	-0.34518	-0.62412	-0.55992	-0.25909	-0.49699	0.00000
Scale-up factors for emission inventories INE99-SINEA05/PROAIRE05/Landa09							
Source	NO_X	SO_2	VOC	CO	PM_{10}	$PM_{2.5}$	NH_3
Area	0.26081	0.13190	0.07201	0.73229	1.93044	1.38817	2.97866
Mobile	3.19318	-0.18559	3.85751	7.71569	-0.48348	-0.65940	1.58809
Non-road mobiles	-0.05984	-0.06241	0.02695	-0.03828	-0.02860	-0.02901	0.00000
Fixed	0.27546	-0.54269	-0.23581	-0.48692	-0.72079	-0.87133	0.00000
Scale-up factors for emission inventories INE99-Landa12							
Source	NO_X	SO_2	VOC	CO	PM_{10}	$PM_{2.5}$	NH_3
Area	0.14828	0.13426	0.01278	3.07269	3.50322	3.50359	-0.96877
Mobile	3.39169	-0.14615	4.39186	8.66790	-0.46290	-0.64661	1.73148
Non-road mobiles*	-0.05984	-0.06241	0.02695	-0.03828	-0.02860	-0.02901	0.00000
Fixed	0.68082	0.41253	-0.29536	2.47504	0.36597	0.29258	0.00000
Scale-up factors for emission inventories INE99-INE05							
Source	NO_X	SO_2	VOC	CO	PM_{10}	$PM_{2.5}$	NH_3
Area	0.07034	0.12787	0.03444	0.62251	1.54565	1.11254	0.13098
Mobile	3.39170	-0.14626	4.39187	8.66790	-0.46287	-0.64669	1.73184
Non-road mobiles*	-0.05984	-0.06241	0.02695	-0.03828	-0.02860	-0.02901	0.00000
Fixed	0.36859	-0.53214	-0.30052	2.43467	-0.04080	-0.10949	0.00000

2.5.2 Meteorological model and parameterization

Meteorological parameters such as wind speed and direction, temperature, humidity, pressure and solar radiation were estimated using the fifth-generation mesoscale weather model (MM5) version 3.7 developed by Pennsylvania State University (PSU) and the National Center for Atmospheric Research (NCAR) (Grell et al., 1994). Details on the configuration of physical options and parameterization used can be found in Vanoye and Mendoza (2009) and Sierra et al. (2013).

2.5.3 Chemical transport model

In the present application, CMAQ version 4.7.1 (Byun and Ching, 1999) chemistry and transport module was chosen for the simulation of chemical transformation and fate of pollutants. CMAQ is a Eulerian model that simulates complex interactions between different air pollutants at regional and urban scales (Dennis et al., 1996). CMAQ can model horizontal advection processes, vertical advection, mass conservation settings for advection processes, horizontal diffusion, vertical diffusion, gas phase reaction, gas-particle transformation processes, photocatalytic reaction calculation, among others. In this application, CMAQ simulated the transport and chemical transformation of tropospheric ozone (O_3) and its precursors for a modeling domain centered in the MMA during the period from 22 to 26 August 2005 (Figure 2.4). O_3 was selected because it is a secondary pollutant, produced in non-linear processes, which makes it a good proxy for overall model performance evaluation.

One of the most important components of air quality models is the photochemical mechanism that describes how volatile organic compounds (VOCs) and nitrogen oxides (NO_x) interact to produce O_3 and other oxidizing species (Yu et al., 2010). In this application, the chemical mechanism Carbon Bond (CB05) was used (Yarwood et al., 2005). CB05 characterizes for employing a "clumping" structure to condense the reactions of individual VOCs. CB05 is an updated version of the CB4 mechanism and includes 59 species and 156 reactions, with updated reaction constants, additional inorganic reactions, and a greater number of organic species than the CB4 version (Yu et al., 2010). Among the chemical species that CB05 can

treat are ethane, olefins, internal, terpenes, formic acid, acetic acid, methanol, ethanol, peroxyacetic acid. Although there are alternative chemical mechanisms such as SAPRC99 -which contains 80 species and 214 reactions-, studies have been conducted comparing its performance in ozone prediction finding that no particle mechanism -between CB5, CB4 and SAPRC99- performs systematically better than another (Yu et al., 2010). Therefore, CB05 was selected, in addition to being the mechanism used by the Environmental Protection Agency (EPA) of the United States in its modeling platform. A major assumption is that the uncertainty provided by the chemical mechanism is negligible.

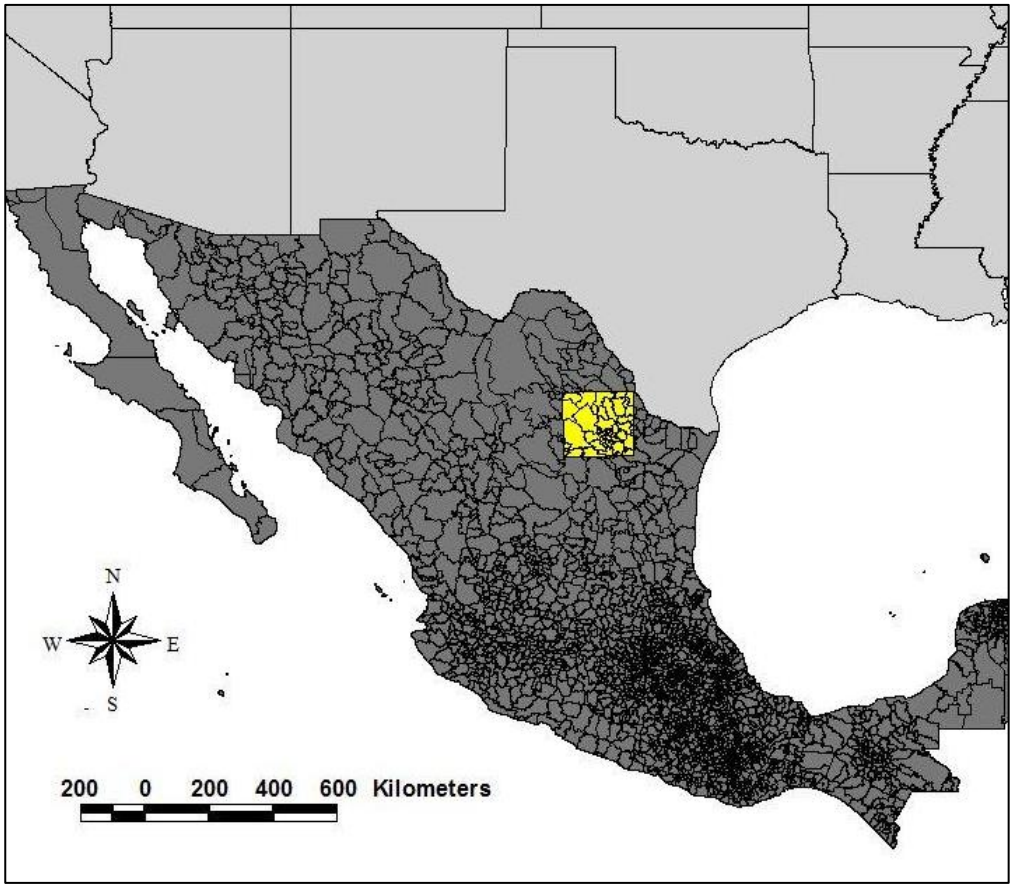

Figure 2.4. Location of the modeling domain centered in the Metropolitan Area of Monterrey.

2.6 Air quality model performance of different emission inventories

The SMOKE-processed inventories were fed into the CMAQ program, keeping the rest of the execution variables as in Sierra et al. (2013). Figure 2.5 presents the dispersion plots for the observed ozone concentrations (computed as the average concentration of O_3 of the 5 monitoring sites within an 8 km x 8 km cell containing the MMA) with respect to the concentrations simulated by CMAQ using the different scaled inventories. The best model performance was obtained with the INE99-INE05 emission inventory, with a R^2 = 0.52, in contrast to other emission inventories with R^2 ranging from 0.41 (INE99-PROAIRE05) to 0.49 (INE99-SINEA05 /PROAIRE05/Landa05). It can also be noted that the presented simulations tend to slightly overestimate the observed concentrations.

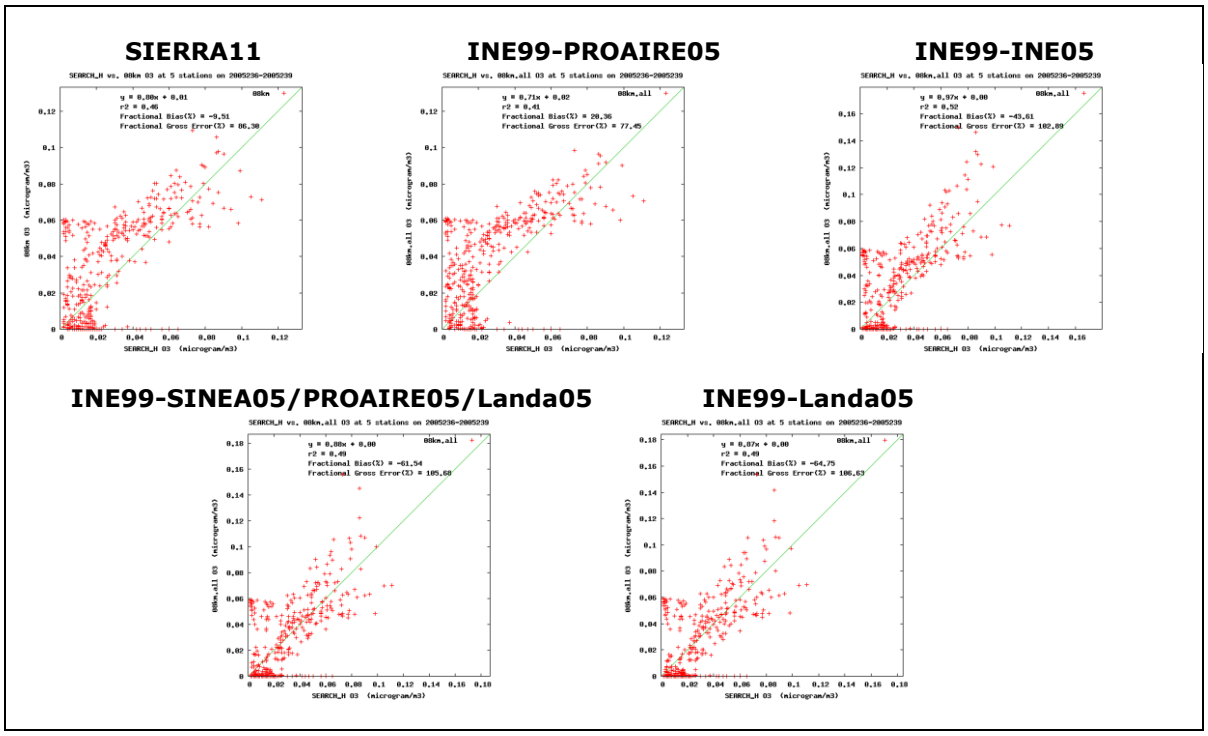

Figure 2.5 Observation dispersion plots vs simulation for the CMAQ photochemical model using different emission inventories.

The differences can be attributed to the use of different methodologies, for example, mobile emissions can be estimated through top-down approaches, based

on state-wide sales of fuels (e.g. gasoline and diesel); or through bottom-up approaches involving the application of models such as MOBILE, a software developed model by the US EPA for estimating pollution from highway vehicles. Another possible source of discordance is source classification. However, while differences are evident, it is difficult to pinpoint a major source of uncertainty as Mexican published inventories do not usually detail all procedures.

2.7 Analysis of available inventories for base years from 2013 to 2018.

The Government of Mexico updated the National Emissions Inventory 1999-2005 for base years 2008, 2013 and 2016, from now on referred to as INEM08, INEM13 and INEM16 emission inventories. On the other hand, the Government of the State of Nuevo León published in September 2016 the Program to Improve Air Quality of the State of Nuevo León 2016-2025 (ProAire 2016-2025), which included an update of the emissions inventory for base year 2013 (referred as PROAIRE13). Tables 2.5 to 2.8 show the estimated emissions, by category, for each of the different criteria pollutant inventories available for the territory of Nuevo León. The inventories presented in Tables 2.5 to 2.8 include the inventories listed in Table 2.1, but complemented with INEM13, INEM16 and PROAIRE13 information.

Table 2.5 Emission inventories for mobile sources (including on-road and non-road), in tons/year.

Coverage	State of Nuevo Leon					
Inventory	INEM99	SINEA05	LANDA05	INEM13	INEM16	PROAIRE13
NO_x	54,636	177,206	177,206	62,104	69,473	37,857
SO_2	2,335	1,831	1,831	2,124	271	436
VOC	54,009	282,848	282,848	12,720	22,109	27,267
CO	398,760	3,783,994	3,783,994	187,987	209,941	287,239
PM_{10}	3,482	969	969	2,654	1,862	836.8
$PM_{2.5}$	3,280	584	584	2,546	1,704	787.7
NH_3	621	1,696	1,696	871	426	1,093

Table 2.6 Emission inventories for area sources in Nuevo León, in tons/year.

Coverage	State of Nuevo Léon					
Inventory	INEM99	SINEA05	LANDA05	INEM13	INEM16	PROAIRE13
NO_x	7,699	8,167	8,840	4403	1,904	4,061
SO_2	17,357	19,576	19,688	233	56	176
VOC	72,793	75,266	73,723	73937	56,535	69,088
CO	25,929	41,560	105,601	17460	11,575	9,984
PM_{10}	5,334	13,532	24,021	5984	8,105	16,476
$PM_{2.5}$	3,558	7,473	16,025	3419	2,524	4,423
NH_3	24,847	28,091	776	18436	18,350	19,146

Table 2.7 Emission inventories for stationary sources in Nuevo León, in tons/year.

Coverage	State of Nuevo León				
Inventory	INEM99	LANDA05	INEM13	INEM16	PROAIRE13
NO_x	22,647	38,065	30,020	30,952	19,619
SO_2	90,401	127,694	23,040	16,679	36,640
VOC	24,624	17,351	9,291	8,256	7,998
CO	24,380	84,720	8,875	27,552	8,366
PM_{10}	11,741	16,038	7,911	9,791	7,793
$PM_{2.5}$	10,386	13,425	5,861	7,208	6,055
NH_3	Nd	776	388	196.9262	172.3

Table 2.8 Comparison of total emissions (all sources) in tons/year.

Coverage	State of Nuevo León			
Inventory	INEM99	INEM13	INEM16	PROAIRE13
NO_x	84,982	96,527	102,329	61,537
SO_2	110,093	25,397	17,006	37,252
VOC	151,426	95,948	86,900	104,354
CO	449,069	214,322	249,068	305,589
PM_{10}	20,557	16,549	19,757	25,105
$PM_{2.5}$	17,224	11,825	11,436	11,266
NH_3	25,468	19,695	18,973	20,412

As an example, Table 2.9 presents the percentage differences between the estimates of the INEM13 vs INEM16, as well as the INEM13 vs PROAIRE13, finding differences of up to 51.7% for PM_{10}, 46.7% for SO_2 and 42.6% for CO emissions published in PROAIRE13 with respect to those reported in the INEM13. On the other hand, PROAIRE13 reports 36.2% less NO_x emissions than INEM13; while, according to federal data, NO_x emissions followed an increasing trend, and grew 6% in the period 2013 to 2016.

Table 2.9 Percentage differences between INEM vs PROAIRE inventories.

Pollutant	NOx	SO2	VOC	CO	PM10	PM25	NH3
INEM13vs16	6.0%	-33.0%	-9.4%	16.2%	19.4%	-3.3%	-3.7%
INEM13vsPRO13	-36.2%	46.7%	8.8%	42.6%	51.7%	-4.7%	3.6%

To explain the observed differences, the relevant activity data needs to be examined. For example, according to the National Institute of Statistics and Geography (INEGI, 2022), the number of vehicles in Nuevo León decreased 13.1% in 2016 with respect to 2013, which might explain part of the decreasing VOC emissions in INEM16 when compared to INEM13 but would not solely explain the increased CO and NO_x emissions.

Moreover, although specific economic sectors exhibited different behaviors in the period 2013-2016, overall, the State of Nuevo Leon showed sustained economic growth during the same period, as measured by its Gross Domestic Product (GDP) (Secretaría de Economía del Estado de Nuevo León, 2022). In the case of INEM13 vs PROAIRE13, the differences must arise from use of different estimation methodologies, source classification, and/or different databases as they relate to the same base year.

In January 2022, the Government of the State of Nuevo León presented, in conjunction with the Nuevo León Council for Strategic Planning (Consejo Nuevo León), the Monterrey Metropolitan Environmental Fund (Fondo Ambiental

Metropolitano de Monterrey), and the Clean Air Institute, an emissions inventory covering the 18 municipalities of the MMA, and base year 2018. Although the absolute amounts by type of pollutant were not publicly reported, a summary table was published with the percentage contribution of each type of emission source to the total emission of each pollutant (Fondo Ambiental Metropolitano de Monterrey, 2022). This information is presented in Table 2.10 under columns labeled NL2018. For comparison, columns INEM16 and PROAIRE13 were added, with their respective contributions by type of source and pollutant.

Table 2.10 Percentage contributions of pollutants in emission inventories.

Pollutant	NL2018			INEM16			PROAIRE13		
	Stationary	Mobile	Area	Stationary	Mobile	Area	Stationary	Mobile	Area
NOx	44%	53%	3%	30%	68%	2%	32%	62%	7%
SO2	89%	10%	1%	98%	2%	0%	98%	1%	0%
VOC	29%	4%	68%	10%	25%	65%	8%	26%	66%
CO	9%	87%	4%	11%	84%	5%	3%	94%	3%
PM10	57%	18%	25%	50%	9%	41%	31%	3%	66%
PM25	65%	14%	20%	63%	15%	22%	54%	7%	39%
NH3	5%	6%	89%	1%	2%	97%	1%	5%	94%

The largest differences between the contributions reported in the different inventories are observed for PM_{10} from stationary sources, which according to the PROAIRE13, in 2013, stationary sources contributed with 31% of PM_{10} emissions while -according to NL2018- in 2018, 57% of PM_{10} emissions came from stationary sources. Accordingly, area sources contributed with 66% of PM_{10} emissions in 2013 (PROAIRE13) but only 25% in 2018 (NL2018). The contribution of stationary sources and mobile sources to VOC emissions also shows a difference between the inventories analyzed, varying between 8% and 29%, and 4% and 26% respectively.

2.8 Conclusions

The analysis presented in this chapter demonstrates that emission inventories can exhibit large uncertainties, hindering the ability of air quality models to accurately predict pollutant concentrations and other intended applications. These uncertainties are mainly derived from the lack of *i*) bottom-up information -which can be very complex and costly to acquire-, *ii*) standardized methodologies, and *iii*) capacity building.

The existence of uncertainties in emission inventories remains a current issue, as exemplified by experiences documented in literature, and is especially relevant for emission inventories developed in the MMA and the State of Nuevo Leon during years 2005 to 2019. For example, emission inventories with base year 2005, presented differences up to +600% in the case of gaseous mobile emissions, among other discrepancies. More recent emission inventories prepared by different government agencies still showed differences ranging from -3.6% to +51.7% for total emissions of specific pollutants, as in the case of the INEM13 and PROAIRE13 PM_{10} estimates.

Differences also emerged when comparing source contributions to total emissions of specific pollutants, even in recent inventories (e.g., NL2018, INEM16 and PROAIRE13). This can have important implications in the development of public policies for pollution abatement, because based on the information presented by a sole inventory, policy makers could pinpoint a certain activity as a major polluter and disregard other emission sources, when in reality this information is subject to a large degree of uncertainty.

Therefore, evidence shows that techniques and strategies to reduce emission inventories must be assessed. These techniques might include mathematical approaches, data-driven techniques such as regularization.

2.9 References

Albert, D. R. (2020). Monte Carlo uncertainty propagation with the NIST uncertainty machine.

Byun, D., & Schere, K. L. (2006). Review of the governing equations, computational algorithms, and other components of the Models-3 Community Multiscale Air Quality (CMAQ) modeling system.

Fondo Ambiental Metropolitano de Monterrey (2022, January 17). *Inventario de Emisiones Atmosféricas del Área Metropolitana de Monterrey*. Retrieved November 1, 2022, from https://famm.mx/archivos/inventario_emisiones_amm.pdf

Gouveia, N., Kephart, J. L., Dronova, I., McClure, L., Granados, J. T., Betancourt, R. M., ... & Diez-Roux, A. V. (2021). Ambient fine particulate matter in Latin American cities: Levels, population exposure, and associated urban factors. *Science of the Total Environment*, *772*, 145035.

Guevara, M., Lopez-Aparicio, S., Cuvelier, C., Tarrason, L., Clappier, A., & Thunis, P. (2017). A benchmarking tool to screen and compare bottom-up and top-down atmospheric emission inventories. *Air Quality, Atmosphere & Health*, *10*(5), 627-642.

Harley R., Russell A.G., McRae G.J., Cass G.R. y Seinfeld J.H. (1993). Photochemical modeling of the Southern California Air Quality Study. Environ. Sci. Technol. 27, 378–388.

Houyoux, M. R., & Vukovich, J. M. (1999). Updates to the Sparse Matrix Operator Kernel Emissions (SMOKE) modeling system and integration with Models-3. *The Emission Inventory: Regional Strategies for the Future*, *1461*, 1-11.

INECC, Instituto Nacional de Ecología y Cambio Climático (2018) México: Inventario Nacional de Emisiones de Gases y Compuesto de Efecto Invernadero 1990–2015 (INEGyCEI).

INEGI, Instituto Nacional de Estadística y Geografía (2022) México: Vehículos de motor registrados en circulación. Available at: https://www.inegi.org.mx/app/tabulados/interactivos/?px=VMRC_2&bd=VMRC. Last accessed: November 30, 2022.

Kuhns, H., Knipping, E. M., & Vukovich, J. M. (2005). Development of a United States–Mexico emissions inventory for the big bend regional aerosol and visibility observational (BRAVO) study. *Journal of the Air & Waste Management Association*, *55*(5), 677-692.

Mendoza-Dominguez, A., Wilkinson, J. G., Yang, Y. J., & Russell, A. G. (2000). Modeling and direct sensitivity analysis of biogenic emissions impacts on regional ozone formation in the Mexico-US border area. Journal of the Air & Waste Management Association, 50(1), 21-31.

Pouliot, G., van der Gon, H. A. D., Kuenen, J., Zhang, J., Moran, M. D., & Makar, P. A. (2015). Analysis of the emission inventories and model-ready emission datasets of Europe and North America for phase 2 of the AQMEII project. *Atmospheric Environment*, *115*, 345-360.

Russell, A. G., McCue, K. F., & Cass, G. R. (1988). Mathematical modeling of the formation of nitrogen-containing air pollutants. 1. Evaluation of an Eulerian photochemical model. *Environmental science & technology*, *22*(3), 263-271.

Scarpelli, T. R., Jacob, D. J., Villasana, C. A. O., Hernández, I. F. R., Moreno, P. R. C., Alfaro, E. A. C., ... & Zavala-Araiza, D. (2020). A gridded inventory of anthropogenic methane emissions from Mexico based on Mexico's national inventory of greenhouse gases and compounds. *Environmental Research Letters*, *15*(10), 105015.

Secretaría de Economía del Estado de Nuevo León (2022) N.L. Crecimiento del PIB por Actividad Económica. DATA Nuevo León. Available at: http://datos.nl.gob.mx/crecimiento-del-pib-por-actividad-economica/. Last accessed on : November 30, 2022.

SEMARNAT, Secretaría de Medio Ambiente y Recursos Naturales (2006). Inventario Nacional de Emisiones de México, 1999.

Sierra Cerda, A. (2011) Sensibilidad de los niveles de ozono a sus precursores en el área metropolitana de Monterrey y el noreste de México. Tesis de Maestría. Tecnológico de Monterrey. México.

Sierra, A., Vanoye, A. Y., & Mendoza, A. (2013). Ozone sensitivity to its precursor emissions in northeastern Mexico for a summer air pollution episode. *Journal of the Air & Waste Management Association*, *63*(10), 1221-1233.

Streets, D. G., Gupta, S., Waldhoff, S. T., Wang, M. Q., Bond, T. C., & Yiyun, B. (2001). Black carbon emissions in China. *Atmospheric environment*, *35*(25), 4281-4296.

Super, I., Dellaert, S. N., Visschedijk, A. J., & Denier van der Gon, H. A. (2020). Uncertainty analysis of a European high-resolution emission inventory of CO 2 and CO to support inverse modelling and network design. *Atmospheric Chemistry and Physics*, *20*(3), 1795-1816.

Vanoye, A. Y., & Mendoza, A. (2009). Mesoscale meteorological simulations of summer ozone episodes in Mexicali and Monterrey, Mexico: analysis of model sensitivity to grid resolution and parameterization schemes. *Water, Air, & Soil Pollution: Focus*, *9*(3), 185-202.

Vivanco Moreno, S. F., & Ramírez Lara, E. (2007). Aplicación del Modelo Hysplit (Hybrid Single Particle Lagrangian Integrated Trayectories) para Evaluar las Trayectorias del Aire y su Impacto en la Dispersión de Contaminantes Atmosféricos.

Zhao, Y., Nielsen, C. P., Lei, Y., McElroy, M. B., & Hao, J. (2011). Quantifying the uncertainties of a bottom-up emission inventory of anthropogenic atmospheric pollutants in China. *Atmospheric Chemistry and Physics*, *11*(5), 2295-2308.

This page intentionally left blank

Chapter 3

Review of Direct Regularization Methods and their Applications in Atmospheric Sciences

This is the English translation of the extended abstract "Revisión de Métodos de Regularización Directa y sus Aplicaciones en las Ciencias Atmosféricas" by Ana Yael Vanoye García and Alberto Mendoza Domínguez, as presented in the L Convención Nacional del Instituto Mexicano de Ingenieros Químicos, during June 2010, in Monterrey, Mexico. The motivation of this work arises from the need to improve emission inventories for air quality modeling applications described in Chapter 2.

3.1 Abstract

Regularization is a mathematical technique for providing numerical stability to an ill-posed problem via the addition of a penalizing term to its formulation. Here an overview of some regularization techniques is presented, with special focus on direct regularization methods, such as Tikhonov regularization, ridge regression, and singular value decomposition. A common feature shared by these regularization techniques is that all of them require the use of a regularization parameter, which controls the weight given to the penalizing term, while seeking a balance between the minimization and regularization errors. Therefore, a description of some of the main methods for choosing this regularization parameter is also presented, including the discrepancy principle, generalized cross-validation, the L-curve method, and normalized cumulative periodogram, as well as some hybrid methods. Regularization with restrictions is also briefly discussed. Finally, typical applications of regularization methods in atmospheric sciences are listed, emphasizing the need to exploit the techniques discussed here in applications where emission strengths are reconstructed or corrected through inverse modeling.

Key words: regularization, Tikhonov, regularization parameter, atmospheric model.

3.2 Introduction

It is common in science and engineering to come across “inverse problems,” where the objective is to determine the value of an unknown parameter by way of measurements (experimental data) indirectly related to the parameter. Thus, while a “direct problem” would be represented as:

$$\mathbf{b}^{\mathbf{prediction}} = \mathbf{A}\mathbf{x}^{\mathbf{measurement}} \tag{3.1}$$

where b is the prediction, x is the observation vector, and A is the physical model or mathematical function that relates observations b with parameter x, an inverse model would be represented as:

$$\mathbf{x}^{\mathbf{prediction}} = \mathbf{A}^{-1}\mathbf{b}^{\mathbf{measurement}} \tag{3.2}$$

Nonetheless, in resolving an inverse problem it is possible for a small amount of error in the measurements or rounding thereof to produce an enormous error in the estimates. This phenomenon of instability in the solution –known as “poor conditioning”– is extremely inconvenient given that, in practice, all measurements taken contain some degree of error, which often is not directly quantifiable. In addition to stability, a well-conditioned linear problem must also meet the conditions of existence and uniqueness (Menke et al., 1989).

Mathematical techniques have been developed for solving poorly conditioned problems. These techniques, based on the incorporation of known properties of the solution x, improve the conditioning of the inverse problem and tend to provide better solutions. Three methods for incorporating information are: *i)* definition of an initial solution, *ii*) relative weighting of the terms, and *iii*) regularization techniques. Moreover, these methods can be combined among themselves (Santamarina and Fratta, 1998). This document focuses on the description of certain regularization techniques, in which regularization means that some restrictions shall be incorporated during the process of building interpolation function $\mathbf{A}$ which, given $\mathbf{x}$,

best fits $\mathbf{b}$, with the aim of stabilizing the problem, reducing the generalization error and finding a stable, physically plausible solution to the inverse problem.

The regularization of an inverse problem is carried out through the addition of a functional (term) that restricts the weighting of the components of the approximation function in accordance with an *a priori* distribution of the weighting, where the regularization matrix is that which imposes such a distribution of the weighting. The regularization technique can be applied with different methodologies for function approximation. However, in almost all of them the magnitude of the regularization depends on regularization parameter λ, which attempts to establish a balance between fitting the desired function to the data set and fulfilling the restrictions imposed by the penalizing functional. The selection of the regularization parameter is a crucial step; many methods have been developed for this purpose. Nonetheless, today there is still a need for efficient and reliable methods for selecting this parameter (Krawczyk-StanDo and Rudnicki, 2007).

The present work provides a general overview of the existing regularization methodologies, with special emphasis on direct methods, such as Tikhonov regularization, ridge regression, truncated singular value decomposition (TSVD), and reduced singular value decomposition. Subsequently, a brief compendium is made of the most commonly used regularization parameter selection methods: discrepancy principle, unbiased predictive risk estimator, generalized cross-validation, the L-curve method, the error consistency method, normalized cumulative histograms, and hybrid methods. The concept of regularization with restrictions is also presented and a few efforts on its use are described. Finally, some general applications on the use of direct regularization methods within the field of atmospheric sciences are listed. Special emphasis is set on the increasing need to use of the techniques described here when inverse-derived emissions are estimated coupling chemical-transport models and inverse modeling techniques.

3.3 Regularization Methods

Regularization is a versatile and useful methodology capable of providing adequate solutions even when faced with errors in the data and model, and which is widely applicable when the unknown solution x refers to discreet values which vary when in three dimensions, in a plane or in a line. A variety of regularization methods have been developed for solving poorly- conditioned problems, which are classified as direct and indirect methods. The direct methods comprise procedures where the estimated solution can be calculated in a single step (which is not exempt from the involvement of some iteration to find the root of a given equation). A characteristic common to all the direct methods is that all of them require a regularization parameter.

The indirect regularization methods tend to be based on the self-regularization property of the specific solution method. The iterative methods possess this property in that the anticipated end of the iterative process has a regularizing effect, as can be seen, for example, in Landweberts iteration formula (Binder et al. 2002), where the iteration index constitutes the regularization parameter. These methods exhibit semi-convergence, which means that the solution improves during the first iterations but deteriorates due to error in the later stages. The iterative regularization methods represent an alternative when methods such as Tikhonov and TSVD do not perform adequately, as sometimes happens with large-scale problems (parameters of ~10^6).

An example of large-scale systems would be the numerical models for weather prediction, which may have as much as 10^7 components in the model vector and 10^5 components in the state vector, resulting in large-scale matrices (Bouttier, 1997). Some iterative methods include: the minimum residual method and its variants (Saad and Schultz, 1986; Jensen and Hansen, 2007), the iteratively regularized Gauss-Newton method (which basically consists of a regularization of Tikhonov with a variable regularization parameter) which has been proven to be

computationally efficient (Binder et al. 2002), the Landwebert method and the Levenberg-Marquardt method (Binder et al. 2002, Doicu et al. 2004).

One of the principal differences between the direct and indirect methods is that while the direct methods demand the explicit calculation of the matrix $\bar{\mathbf{A}} = \mathbf{A}\mathbf{L}^{-1}$ by means of standard methods such as QR factorization, iterative methods only require the efficient calculation of the product $\bar{\mathbf{A}}\bar{\mathbf{x}}$ (Hansen, 1994), which is simpler from a numerical perspective. Nonetheless, one of the principal disadvantages of the indirect or iterative methods lies in the difficulty of establishing the point at which the iteration should end. The regularly used methods for well-conditioned iterative problems, such as those based on the residual, tend not to work adequately for poorly- conditioned problems. Moreover, an imprecise estimation of the iteration´s end point can produce a solution with a relative error that is significantly greater than that of the optimal solution (Chung and Nagy, 2010).

3.4 Direct Regularization Methods

Following is a more detailed description of some commonly used direct regularization methods.

3.4.1 Tikhonov Regularization

The most common and frequently used regularization method is Tikhonov regularization (Willoughby, 1979; Neumaier, 1998). Tikhonov regularization is a powerful tool for solving poorly-conditioned linear systems and for linear least squares problems. It consists of the substitution of 2-norm minimization problem, which corresponds to the least squares problem:

$$\min_{u \in C^n} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2 \qquad (3.3)$$

–where $\mathbf{x}$ is the unknown vector, $\mathbf{b}$ is the observation vector, and $\mathbf{A}$ is an $\boldsymbol{m} \times \boldsymbol{m}$ matrix, with a large condition number and $\boldsymbol{m} \geq \boldsymbol{n}$ ($\boldsymbol{n}$ is the number of parameters

and $\boldsymbol{m}$ is the number of data points)– for a problem that includes a Tikhonov functional:

$$\mathbf{x}_{\boldsymbol{\lambda}} = \mathbf{min}\left\{\|\mathbf{Ax} - \mathbf{b}\|_2 \quad + \|\boldsymbol{\Gamma}\mathbf{x}\|_2 \quad \right\} \tag{3.4}$$

In Equation (3.4), $\boldsymbol{\Gamma} = \boldsymbol{\lambda}\mathbf{L}$, $\mathbf{L}$ tends to be the identity matrix and $\lambda \in \mathbb{R}$ represents the regularization parameter. The Tikhonov functional is a measure of the "smoothness" of solution $\mathbf{x}$ and penalizes its discontinuities. In the Tikhonov regularization, the regularized solution $\mathbf{x}_{\boldsymbol{\lambda}}$ results from minimizing the weighted combination of the residual norm and an additional restriction. Regularization parameter λ controls the weight given to the minimization of the additional restriction as relates to the minimization of the residual norm and as was mentioned, the perturbation error and the regularization error must be balanced in the regularized solution. Thus, a large λ (equivalent to a large amount of regularization) favors a semi-norm of the lesser solution at the expense of a large residual norm, while a small λ (little regularization) will have the opposite effect. λ also controls the sensitivity of the regularized solution $\mathbf{x}_{\boldsymbol{\lambda}}$ to perturbations in $\mathbf{A}$ and $\mathbf{b}$.

Lewis et al. (2006) present algebraic proof as to how, under certain conditions, Tikhonov regularization corresponds equally to the solution of the minimum residual as to the minimum norm. The regularized solution to the problem presented in Equation (3.4) will be:

$$\mathbf{x}_{\boldsymbol{\lambda}} = (\mathbf{A}^{\mathbf{T}}\mathbf{A} + \boldsymbol{\Gamma}^{\mathbf{T}}\boldsymbol{\Gamma})^{-\mathbf{1}}\mathbf{A}^{\mathbf{T}}\mathbf{b} \tag{3.5}$$

Nonetheless, in the case that L is equal to the identity matrix, it is also possible to solve the least squares problem using filter factors and the generalized singular value decomposition (GSVD) explicitly, or singular value decomposition (SVD). Let $\mathbf{A}$ be a matrix whose SVD is:

$$\mathbf{A} = \sum_{i=1}^{n} \boldsymbol{\sigma}_i \boldsymbol{u}_i \boldsymbol{v}_i^{T} = \mathbf{S}\,\mathbf{U}\,\mathbf{V}^{\mathbf{T}} \tag{3.6}$$

where the singular left and right vectors $\boldsymbol{u}_i$ and $\boldsymbol{v}_i$, respectively, are orthonormal, while the singular values $\boldsymbol{\sigma}_i$ are non-negative and non-incremental numbers (i.e., $\boldsymbol{\sigma}_1 \geq \boldsymbol{\sigma}_2 \geq \cdots \geq \boldsymbol{\sigma}_n \geq \mathbf{0}$). Using the SVD of A, it is easy to prove that the solution to the ordinary least squares (LSQ) problem denoted in Equation (3.3) can also be represented as:

$$\boldsymbol{x}_{LSQ} = \sum_{i=1}^{n} \frac{\boldsymbol{\alpha}_i}{\boldsymbol{\sigma}_i} \boldsymbol{v}_i = \sum_{i=1}^{n} \frac{\boldsymbol{u}_i^{T} \boldsymbol{b}}{\boldsymbol{\sigma}_i} \boldsymbol{v}_i = \frac{\mathbf{U^T b}}{\mathbf{s}} \mathbf{V} \tag{3.7}$$

The problem of using the least squares solution $\boldsymbol{x}_{LSQ}$ consists in that the error in the directions corresponding to the small singular values ($\boldsymbol{\sigma}_i$) is amply magnified and opaque the information contained in the directions corresponding to the larger singular values (Hansen and O'Leary, 1993). The regularization methods thus incorporate filter factors fi, transforming the solution to:

$$\boldsymbol{x}_{filtro} = \sum_{i=1}^{n} \boldsymbol{f}_i \frac{\boldsymbol{\alpha}_i}{\boldsymbol{\sigma}_i} \boldsymbol{v}_i \tag{3.8}$$

and differ in the selection method of said parameter. For example, a filter factor equal to 1 corresponds to the ordinary least squares solution. The filter factor for Tikhonov regularization consists of:

$$\boldsymbol{f}_i = \frac{\boldsymbol{\sigma}_i^2}{\boldsymbol{\sigma}_i^2 + \boldsymbol{\lambda}^2} \tag{3.9}$$

3.4.2 Ridge Regression

Tikhonov regularization can also be formulated from a statistical perspective, known as ridge regression (Hoerl and Kennard, 1979), which proposes an estimation procedure based on:

$$\mathbf{x}^{*} = \lfloor \mathbf{A}'\mathbf{A} + \mathbf{K} \rfloor^{-1} \mathbf{A}'\mathbf{b} \tag{3.10}$$

Where $\mathbf{x}^*$ is the "regularized" solution and $\mathbf{K}$ is a diagonal matrix of non-negative constants, and analogous to the Tikhonov functional when the $\mathbf{L}$ matrix is equal to the identity:

$$\mathbf{K} = k\mathbf{I}, \quad k \geq 0 \tag{3.11}$$

3.4.3 Truncated Singular Value Decomposition (TSVD)

The Picard condition (Hansen, 1990) ensures the stability of the solution of the inverse problem, and it is met when the dot products of the columns $\mathbf{U}$ and the data vector $\mathbf{b}$ fall to zero faster than the singular values σ_i. Under this condition, no instability will be observed due to small singular values (Aster et al. 2005). Nonetheless, even when the Picard condition is not met, it is possible to recuperate a useful model by truncating the sum in Equation (3.7) to produce a solution through TSVD (Aster et al. 2005). TSVD is a common regularization method which consists of simply truncating the sum in Equation (3.12) in a higher limit $k < n$, before the small singular values begin to dominate (Hansen and O'Leary, 1993). It is possible to note that when $k = n$ (i.e., when the small singular values are included), the solution obtained through the TSVD method is identical to that produced through the ordinary least squares method. Nonetheless, according to the theory, a solution obtained through TSVD where $k < n$ would be more stable (Aster et al. 2005). In terms of the SVD, the filter factor for the TSVD would be between 0 and 1 (Hansen, 2008).

3.4.4 Reduced Singular Value Decomposition

Reduced singular value decomposition (RSVD) (Ekstrom and Roads, 1974) is considered to be a regularization method similar to Tikhonov, in terms of its SVD, with the difference that the filter factor f_i for RSVD is (Chung and Nagy, 2010): $f_i = \frac{\sigma_i}{\sigma_i+\lambda}$ when $\mathbf{L} = \mathbf{I_n}$, and $f_i = \frac{\sigma_i}{\sigma_i+\lambda\mu_i}$, when $\mathbf{L} \neq \mathbf{I_n}$.

3.4.5 Other Direct Regularization Methods

There are other direct regularization methods, such as maximum entropy (Smith and Grandy, 1985), total least squares (Fierro et al. 1997), and the Lagrange

method (Landi, 2008), among others. Kitagawa et al. (2001), for example, proposed a method based on QR factorization which has the advantage of being less computationally demanding than SVD, though it also requires estimating a regularization parameter such as that required in Tikhonov regularization. Furthermore, studies have been done on regularization error estimates (Doicu et al. 2007; Lorenz and Rösch, 2010)

3.4.6 Hybrid Methods

The hybrid methods are an alternative that seeks to attenuate the disadvantages inherent to the direct and indirect methods (e.g., the computational inefficiency of direct methods such as Tikhonov regularization in large-scale applications and the semi-convergence of the iterative methods). Two categories of hybrid methods can be distinguished: those using iterative methods to solve the regularized problem and those that couple the regularization with an iterative scheme (Chung, 2009). An example of this type of method is that proposed by Kilmer and O´Leary (2001), which reprojects the problem in smaller dimensions, to then proceed with the solution.

3.5 Selection of Regularization Parameters

There are algorithms that automatically analyze the problem and select the adequate regularization parameter $\boldsymbol{\lambda}$ based on different criteria. However, there is no systematic evaluation and comparison on how these parameter-selecting algorithms perform in the solution of different problems (Doicu et al. 2010). Ideally, $\boldsymbol{\lambda}$ should establish an adequate balance between precision and resolution and be determined through direct methods (Ceccherini, 2005). The key to successful regularization lies in selecting a parameter that is sufficiently large as to stabilize the inversion with respect to the amplification of the perturbation, but not so large that it dominates the original governing equations of matrix $\mathbf{A}$ with the smoothing matrix of $\boldsymbol{\lambda}$L. The complexity of finding an adequate parameter lies in the difficulty of separating the exact data error along with the physical significance of that error,

which is unknown in both quantity and correlation, as well as the discretization error (Åkesson and Daun, 2008). The different methods for selecting the regularization parameter tend to be classified in two categories: those requiring prior knowledge of the variance of the error and those methods that do not require it. Likewise, most of the methods for regularization parameter selection have been designed for use with the Tikhonov functional, and few efforts have been made for other functionals such as that of total variation (Landi, 2008). A few commonly studied methods are:

3.5.1 Discrepancy Principle

The discrepancy principle (Morozov, 1966) is probably the simplest method for selecting the regularization parameter (Hansen and O'Leary, 1993) and consists solely of defining the parameter at such a value that the residual vector norm is less than or equal to a determined tolerance:

$$\|\mathbf{Ax} - \mathbf{b}\|_2 \leq \boldsymbol{\delta} \tag{3.12}$$

In other words, it is based on the reasoning that the least squares residual must have at least the same order of magnitude as the error, so that the regularization parameter must be as large as possible (Åkesson and Daun, 2008). When the number of parameters n is greater than or equal to the number of data m, there is no distribution $\boldsymbol{\chi^2}$ with a negative number of degrees of freedom. In practice, a common heuristic is to require that $\|\mathbf{Ax} - \mathbf{b}\|_2$be less than $\sqrt{\boldsymbol{m}}$, given that the approximate median of a distribution $\boldsymbol{\chi^2}$ with m degrees of freedom is m (Aster et al. 2005). A characteristic (and possible disadvantage) of the discrepancy principle is that it requires prior knowledge of the error or variance (Kilmer and O'Leary). It is important to note that, as $\boldsymbol{\delta}$ tends towards zero, the regularized solution based on the discrepancy principle is convergent but is based on difficult to obtain or erroneous information. Even with a correct estimate of the variance, the solutions tend to be overly smoothed (Kilmer and O'Leary, 2001).

3.5.2 Unbiased Predictive Risk Estimator

The UPRE method requires some prior knowledge about the error. It is based on the minimization of the predictive error, defined as (Lin and Wohlberg, 2008):

$$\frac{1}{n}\|\mathbf{P}_{\lambda}\|^2 = \frac{1}{n}\|\mathbf{A}\mathbf{x}_{\lambda} - \mathbf{A}\mathbf{x}_{real}\|^2 \qquad \textbf{(3.13)}$$

where $\boldsymbol{x}_{\lambda}$ is the solution calculated for parameter $\boldsymbol{\lambda}$, and $\mathbf{x}_{real}$ is the real solution. Given that $\mathbf{x}_{real}$ is, in practice, unknown, it is necessary to define an estimator to estimate the value of optimal $\boldsymbol{\lambda}_{\mathbf{opt}}$. Vogel (2002) showed that:

$$\mathbf{UPRE}_{Tikh}(\boldsymbol{\lambda}) = \frac{1}{n}\|\mathbf{r}_{\lambda}\|^2 + \frac{2\sigma^2}{n}\mathbf{traza}\,(\mathbf{A}_{\lambda}) - \boldsymbol{\sigma}^2 \qquad \textbf{(3.14)}$$

giving way to:

$$\boldsymbol{\lambda}_{opt} = \mathbf{min}\,\{\mathbf{UPRE}_{Tikh}(\boldsymbol{\lambda})\}\,. \qquad \textbf{(3.15)}$$

Although the computation of $\mathbf{de\ UPRE}_{Tikh}(\boldsymbol{\lambda})$ is relatively straightforward if the SVD of $\mathbf{A}$ is available, said calculation is computationally costly for large-scale problems. In particular, the computation of the trace of the influence matrix is especially demanding, $\mathrm{trace}\{\mathbf{A}(\mathbf{A}^{\mathrm{T}}\mathbf{A}+\lambda\mathbf{I})^{-1}\mathbf{A}^{\mathrm{T}}\}$ (Lin and Wohlberg, 2008). Methods have been developed that seek to solve this problem, such as that developed by Kilmer and O'Leary (2001), which uses Lanczos' procedure to approximate the eigenvalues of the large-scale system´s matrix to a smaller-scale matrix, while Lin and Wohlhberg (2008) combine procedures to calculate the approximate trace, obtaining similar results to those obtained using the exact trace but significantly reducing the computational requirements.

3.5.3 Generalized Cross-validation

Generalized cross-validation (GCV) (Golub et al. 1979; Haber and Oldenburg, 2000) is a technique derived from ordinary cross-validation (OCV), the basic idea of which is as follows: If you eliminate a given observation b_i, and calculate a

solution $\mathbf{x}_{\lambda,i}$, then the estimate of $\boldsymbol{b_i}$ calculated based on $\mathbf{x}_{\lambda,i}$ should thus be a good estimate. While OCV depends on the particular way the data is ordered, GCV is invariant to the orthogonal transformations and permutations of the data vector $\mathbf{b}$ (Hansen and O'Leary, 1993). The GCV function to be minimalized is:

$$\boldsymbol{G(\lambda)} \equiv \frac{\|(\mathbf{Ax}(\boldsymbol{\lambda})-\mathbf{b})\|_2^2}{(\mathbf{traza}\ (\mathbf{I}-\mathbf{AA}(\boldsymbol{\lambda})^{\boldsymbol{I}}))^2} \tag{3.16}$$

where $\mathbf{A}(\boldsymbol{\lambda})^{\boldsymbol{I}}$ is any matrix that maps the $\mathbf{b}$ right-side to the solution $\mathbf{x}(\boldsymbol{\lambda})$, for example, $\mathbf{x}(\boldsymbol{\lambda}) = \mathbf{A}(\boldsymbol{\lambda})^{\boldsymbol{I}}\mathbf{b}$. Upon deriving, it is assumed that the errors on the right side are normally distributed with a median of zero and a covariance matrix of $\boldsymbol{\sigma^2}\mathbf{I}$. An advantage of this method is that it does not require prior knowledge about the error variance (Kilmer and O'Leary). Nonetheless, difficulties may arise when the matrix G possesses a "very flat" minimum, thereby making the numerical determination of parameter λ difficult. Likewise, another inconvenience lies in the fact that sometimes the GCV method can confuse variance correlated with a signal. The method can produce unsatisfactory results when the errors are highly correlated amongst one another. Variations of this technique have been developed, such as the weighted GCV method (Chung, 2009).

3.5.4 L-curve Method

The L-curve method consists in graphing the regularized solution´s norm against the corresponding residual´s norm for each set of regularization parameter values (Hansen and O'Leary, 1993), which in effect results in an L-shaped curve. The L-curve is basically composed of two parts: the more horizontal part corresponds to the solutions where the regularization parameter is very large, and the solution is dominated by the regularization errors, while the vertical part corresponds to the solutions where the regularization parameters are dominated by the approximation errors (Krawczyk-Stando and Rudnicki, 2007).

Thus, the L-curve allows for analysis of how the regularized solution changes as

the regularization parameter changes. The corner of the L-curve can be interpreted as the point closest to the origin or as the point where the curvature is greatest (Hansen and O'Leary, 1993). Hence, the best regularization parameter must be located at the corner of the L-curve, given that for superior values the residual increases without significantly reducing the solution´s norm, while for inferior values, the solution´s norm increases rapidly without a significant change in the residual. In practice, only a certain number of points are evaluated, and the corner is located by calculating the point of greatest curvature (Hansen and O'Leary, 1993; Kilmer and O'Leary, 2001).

Obtaining the L-curve is a numerically manageable method, in addition to it not requiring prior knowledge about the error (Li and Wohlberg, 2008) Nonetheless, some of its limitations consist in that the solution does not achieve convergence with the real solution as n tends toward infinity or as the error´s norm approaches zero (Kilmer and O'Leary, 2001). In fact, all the methods that do not suppose prior knowledge of the error present the latter property. Another difficulty arises when the solution is very "smooth" or "wrinkly," such that the corner may not represent the optimal regularization parameter, such as when, under certain circumstances, the L-curve may simply not have an absolute corner (Åkesson and Daun, 2008).

3.5.5 Error Consistency

This is based on the statement that the regularization must be as strong as possible but should not introduce more smoothing than there is error. Strong regularization is obtained when the smoothing is consistent with the non-regularized profile errors, while regularization is weak when the smoothing is consistent with the regularized profile errors. Ceccherini (2005) showed that the regularization parameter $\boldsymbol{\lambda}$ can be analytically obtained from the non-regularized profile and its variance-covariance matrix:

$$\boldsymbol{\lambda} = \left[\frac{n}{(\mathbf{x_a}-\hat{\mathbf{x}})^{\mathbf{T}}\mathbf{R}\,\mathbf{S}_{\hat{\mathbf{x}}}\,\mathbf{R}\,(\mathbf{x_a}-\hat{\mathbf{x}})}\right]^{1/2} \qquad \textbf{(3.17)}$$

where $\mathbf{x_a}$ is the a priori state vector, $\mathbf{R}$ is the regularization matrix, $\hat{\mathbf{x}}$ is the non-regularized state vector obtained when only the first term of $\mathbf{f(x)}$ is minimized, $\mathbf{S_{\hat{x}}}$ is its variance-covariance matrix and n is the number of elements in the state vector. Due to its analytical formulation, this method can be used in iterative regulation and in operational analysis (Ceccherini, 2005).

3.5.6 Normalized Cumulative Periodogram

The normalized cumulative periodogram (NCP) was first introduced by Rust (1998, 2000) to distinguish the signal from the error in an extension of the truncated SVD algorithm. The regularization parameter selection method for NCP (Rust and O'Leary, 2008) is based on making the magnitude of the residuals come as close as possible to the "white" error, using a diagnostic test based on the periodogram.

Based on the properties of the true residual $\boldsymbol{\eta}$, Rust suggested the following set of diagnostic tests for evaluating the acceptability of a regularized solution $\tilde{\boldsymbol{x}}$ with residual $\tilde{\boldsymbol{c}}$. A formal test of the diagnostics proposed by Rust and O´Leary (2008) is based on the periodogram´s graph, which is an estimate of the potency spectrum of a signal. To build the periodogram, a signal is extended with zeros (zero-padding) at a convenient longitude N (preferably with potency of 2), the Fourier transform is calculated for this augmented series, and squares are taken of the absolute values of the first half of the coefficients. The cumulative periodogram will be the vector of the partial summations of the normalized periodogram by the sum of all the elements.

As a result of the properties of the white noise, a graph of the elements c against their frequencies will be a straight line passing through the origin with slope $2/NT$, where T is the spacing of the samples for the time variable. Finally, although the graph of the residual vector, its periodogram and its cumulative periodogram together with the diagnostic tests provides a good regularization parameter, it is necessary to have numerical criteria that allow for automated selection. In accordance with the established diagnostics, it would be possible to select the

parameter in agreement with some of the rules proposed by Rust and O´Leary (2008).

There is evidence that the use of residual periodogram has produced better results when compared with other standard techniques such as the discrepancy principle, the L-curve, and the GCV (Rust and O'Leary, 2008). Hansen et al. (2007) and Mead (2008) also have used properties of the distribution of the residual for selecting the regularization parameters.

3.5.7 Hybrid Methods

Other hybrid algorithms include those proposed by Frommer and Maass (1999), who sought to approximate λ so as to meet the discrepancy principle by way of focusing on the truncated conjugated gradient, Baglama et al. (1998), who imposed superior and inferior limits to the L-curve using Lanczos´ process of bidiagonalization, and Kaufman and Neumaier (1997), who also used the conjugated gradient for Tikhonov with an L-curve, but maintaining a restriction of non- negativity (Kilmer and O'Leary, 2001). Kilmer and O´Leary (2001) also presented a comparison of the computational requirements for different regularization methods, including Tikhonov and the hybrid projection method. There is also a set of works developed with the purpose of performing a direct comparison between the different regularization parameter selection techniques. These works use the collection of problems presented by Hansen (1998), and among them are those performed by Krawczyk-Stańdo and Rudnicki (2007), Kilmer and O´Leary (2001), Rust and O´Leary (2008), Wu (2003), Viloche-Bazán (2008), and Rezghi and Hosseini (2009).

3.6 Regularization with restrictions

Even when the regularization methods produce more stable and precise solutions, these solutions often lack physical sense or violate some restriction imposed by the nature of the problem. Methodologies have thus been investigated for incorporating

additional restrictions that limit the solution to the problem to a determined set of values or that incorporate some other additional information about the solution.

Many of these methodologies have been developed as special cases of the least squares problem and taking the Tikhonov regularization as a base (Friedrich and Hofmann, 1987). Other techniques for the solution of poorly-conditioned problems with restrictions include the method of penalized maximum likelihood (Bardsley and Goldes, 2009), the Bayesian methods, which are commonly known as stochastic versions of Tikhonov (Hansen, 2008), and the method of maximum entropy (Smith and Grandy, 2013; Wang et al. 2007), which ensures the positivity of the solution. Doicu et al. (2010), Stark and Parker (1995), and Calvetti and Reichel (2004) have also developed relevant works in this regard.

3.7 Application of regularization techniques in atmospheric sciences

As in other fields of science and engineering, poorly-conditioned inverse problems are ubiquitous in atmospheric sciences. Thus, the necessity has arisen to make use of regularization techniques to solve problems such as the reconstruction of vertical profiles of concentrations of atmospheric constituents (Doicu et al., 2007, Doicu et al. 2010, Fedorova et al. 2009; Osterloh et al. 2009, Koner and Drummond, 2008a; Koner and Drummond, 2008b; Ceccherini et al., 2007; Rozanov et al., 2007; Meijer et al, 2006; Ceccherini, 2005; Doicu et al. 2005a; Doicu et al., 2005b; Doicu et al., 2003; Hasekamp and Landgraf, 2001; Englert et al., 2000; Shimpf and Shreier, 1997; Gaikovich, 1994) and other atmospheric parameters (Doicu et al. 2004; Arikan et al., 2003; Zhao et al., 2003; Qin et al., 2002; Eriksson, 2000; Shimpf y Shreier, 1997) based on spectrometric measurements and remote sensing, estimation of particle size distributions of atmospheric aerosols (Cuccia et al., 2010; Wang et al., 2010; Riefler and Wriedt, 2008; Wang, 2008; Wang et al., 2007; Böckmann and Kirsche, 2006; Wang et al., 2006; Lemmety et al., 2005; Talukdar and Swihart, 2003; Voutilainen, 2001; Voutilainen et al., 2000; Wolfenbarger and Seinfeld, 1990), analysis and

reconstruction of climatic series (Horenko, 2010; Christiansen et al. 2009; Drignei et al. 2008), determination of the concentration of chemical species in the environment (Lau et al., 2009; Salcedo-Sanz et al., 2009), location of sensors for monitoring atmospheric contamination (Haber et al. 2008), determination of the global distribution of lightning (Ando and Hayakawa, 2007), inverse modeling of global species cycles (Bergamaschi et al. 2000), estimation of the total vertical content of electrons (Arikan et al., 2003), among other applications.

In solving such problems, diverse regularization techniques have been employed, such as: Tikhonov regularization (e.g., Wang, 2008; Eckhardt et al., 2008; Haber et al., 2008; Nikoukar et al., 2008; Riefler y Wriedt, 2008; Ando and Hayakawa, 2007; Ceccherini, 2005; Lemmety et al, 2005; Krakauer et al, 2004; Talukdar and Swihart, 2003; Zhao et al, 2003; Eriksson, 2000; Goikovich, 1994), Twomey-Tikhonov regularization (Fedorova et al., 2009), Phillips-Twomey regularization (Wang, 2008; Riefler and Wriedt, 2008), Phillips-Tikhonov regularization (Meijer et al., 2006; Hasekamp and Landgraf, 2001; Englert et al., 2000; Shimpf and Shreier), Phillips-Tikhonov-Twomey regularization (Qin et al. 2002), as well as certain other variations thereof (Cuccia et al., 2010; Wang et al., 2010; Doicu et al., 2010; Henze et al., 2009; Wang et al., 2006; Doicu et al., 2003; Wolfenbarger and Seinfeld, 1990), hyperviscous and pseudo-parabolic regularization (Gustafsson and Protas, 2010), ridge regression (Christiansen et al., 2009), and Bayesian regularization (De Vito et al., 2009; Kopacz et al., 2009; Lau et al., 2009; Salcedo-Sanz et al., 2009; Stohl et al., 2009).

Within the atmospheric sciences there is an inverse problem where, in the solution thereof, the regularization techniques described above have yet to be fully employed. This inverse problem includes the inverse modeling of emission sources strengths and the estimation of the distribution of sources and sinks of atmospheric chemical species. Here, typically a chemical-transport model (the “forward” model) is coupled with an inverse modeling technique to derive a corrected emissions field based on the minimization of the error between the model- derived concentrations

and available observations. Although there are many applications reported on the use of inverse modeling to improve emissions inventories, whether globally (e.g. Pétron et al., 2002) or continentally (e.g., Elbern et al., 2007), or even locally (e.g., Quélo et al., 2005), and with the exception of some particular works, the use of regularization techniques for the optimization of such inventories has yet to be fully exploited.

The “top-down” analysis of emissions inventories has been formulated (and attempted to be solved) as an inverse problem for the estimation of chemically unreactive (or relatively unreactive) or reactive species. There exists a vast number of applications of such diversity, including the estimation of emissions of stratospheric ozone depletion substances (e.g., Xiao et al., 2010), greenhouse gases (e.g., Stohl et al., 2009), tropospheric ozone precursors (e.g., Napelenok et al., 2008), aerosol precursors (e.g., Gilliland et al., 2006), specific biogenic species (e.g. Chang et al., 1996), and other atmospheric trace gases such as CO (e.g. Pétron et al., 2002).

In these efforts, two foci have traditionally been used for optimizing such emission inventories: the assimilation of data through Kalman or ensemble Kalman filtering (KF or EnKF) and 4-D variational assimilation (4-DVar). Despite their frequent application, these methods present the disadvantages of, in the case of 4-DVar, requiring the computation of the adjoint model and the objective function gradient (which are difficult to obtain for complex three- dimensional chemistry-transport models), or erroneously supposing a Gaussian error distribution in the case of KF or EnKF (Kalnay et al., 2007). Li et al. (2010) has also used genetic algorithms for optimizing inventories, but their application is still limited due to the necessary computational requirements.

Fewer examples exist on the use of regularization techniques to obtain inverse-derived emissions inventories. In particular, few formal methods have been applied to obtain the value of the regularization parameters. For example, Mendoza-Dominguez and Russell (2000, 2001) used ridge regression to deduce, separately,

the emissions of anthropogenic and biogenic tropospheric ozone precursors. In their proposed method, a stepwise approach to calculate matrix K in equation (10) is taken. First, the ridge parameter λ is estimated based on the work of the Hoerl and Kennard (1976). Then, a length parameter (l), which expands or shrinks the ridge regression solution vector, is computed based on the work of Aldrin (1997). If the solution obtained by the application of the above process is beyond what is expected to be a plausible solution, Mendoza-Dominguez and Russell (2000) apply differentiated further penalization to each of the diagonal elements of K. This same technique was further explored to correct emissions inventories of primary aerosols and ozone precursors simultaneously (Mendoza-Dominguez and Russel, 2005).

In the same way, Henze et al. (2009) used the L-curve technique and an analysis of the total error and its components to establish the best value of α in their application, Chai et al. (2009) used TSVD, Saide et al. (2009) defined α as the ratio of the mean value of the error variances of the observations and the parameters and obtained it through the L-curve method, Krakauer et al. (2009) used GCV, and Fan et al. (1999) compared the use of three different regularization techniques, including Bayesian inversion and TSVD.

Even though others have acknowledged the value of regularization in attempting to obtain inverse-derived emission estimates, some still follow the strategy of assigning values to the regularization parameter subjectively or empirically. This includes the case of 4-DVar applications where in the cost function to be minimized a regularization parameter can be stated. For example, Eckhardt et al. (2008) used Tikhonov regularization to estimate SO_2 emissions from volcanic eruptions. However, the weight of the smoothness condition “was determined subjectively as ten times the average standard error of the a priori values”. In this context, the application of regularization techniques and formal establishment of the regularization parameter in this type of problems is worth investigating more deeply.

3.8 Conclusions

Regularization is a mathematical technique that offers numerical stability to a poorly-conditioned problem through the addition of a penalizing functional is its formulation. Regularization techniques differ among themselves according to the mathematical form of said functional and the quantity of information to be incorporated, and can be classified in direct, indirect or hybrid. A common characteristic among the direct regularization techniques (of which Tikhonov regularization is the most widely used) is that they all require a regularization parameter, which seeks to balance the minimization error and the regularization error. Among the most frequently used techniques for the selection of this parameter are the discrepancy principle, the L-curve, GCV and NCP, which have produced positive results when compared to the traditional methods. Hybrid methods and regularization methods with restrictions have also been developed.

The regularization methods described here are applicable in multiple fields of the atmospheric sciences, including the estimation of emissions inventories. In this last, inverse modeling techniques have favored the use of Kalman or Ensemble Kalman Filtering, or 4-D variational assimilation, though some efforts have been documented on the use of ridge regression or Tikhonov regularization. The methodologies that have been derived to determine the best regularization parameter can be extended to the practice of obtaining inverse- derived emission strengths of atmospheric constituents.

In the following chapters, some regularization methods will be used along a deterministic photochemical air quality model to compute scaling factors for the correction of a criteria-pollutant emission inventory for Guadalajara Metropolitan Area, Mexico. The regularization methods to be examined are Tikhonov Regularization (TIKH), TSVD, and DSVD, in combination with the regularization parameter selection methods GCV, LC, and NCP, along a Bounded Variable Least Squares (BVLS) method.

3.9 References

Åkesson, E. O., & Daun, K. J. (2008). Parameter selection methods for axisymmetric flame tomography through Tikhonov regularization. *Applied optics*, *47*(3), 407-416.

Aldrin, M. (1997). Length modified ridge regression. *Computational statistics & data analysis*, *25*(4), 377-398.

Ando, Y., & Hayakawa, M. (2007). Use of generalized cross validation for identification of global lightning distribution by using Schumann resonances. *Radio Science*, *42*(02), 1-7.

Arikan, F. E. Z. A., Erol, C. B., & Arikan, O. (2003). Regularized estimation of vertical total electron content from Global Positioning System data. *Journal of Geophysical Research: Space Physics*, *108*(A12).

Aster, R. C., Borchers, B., & Thurber, C. H. (2005). *Parameter estimation and inverse problems*. Elsevier.

Baglama, J., Calvetti, D., Golub, G. H., & Reichel, L. (1998). Adaptively preconditioned GMRES algorithms. *SIAM Journal on Scientific Computing*, *20*(1), 243-269.

Bardsley, J. M., & Goldes, J. (2009). Regularization parameter selection methods for ill-posed Poisson maximum likelihood estimation. *Inverse Problems*, *25*(9), 095005.

Bazán, F. S. V. (2008). Fixed-point iterations in determining the Tikhonov regularization parameter. *Inverse Problems*, *24*(3), 035001.

Bergamaschi, P., Hein, R., Heimann, M., & Crutzen, P. J. (2000). Inverse modeling of the global CO cycle: 1. Inversion of CO mixing ratios. *Journal of Geophysical Research: Atmospheres*, *105*(D2), 1909-1927.

Binder, T., Blank, L., Dahmen, W., & Marquardt, W. (2002). On the regularization of dynamic data reconciliation problems. *Journal of Process Control*, *12*(4), 557-567.

Böckmann, C., & Kirsche, A. (2006). Iterative regularization method for lidar remote sensing. *Computer Physics Communications*, *174*(8), 607-615.

Bouttier, F. (1997) Meteorological Training Course Lecture Series: Kalman Filtering. European Centre for Medium-Range Weather Forecasts.

Calvetti, D., & Reichel, L. (2004). Tikhonov regularization with a solution constraint. *SIAM Journal on Scientific Computing*, *26*(1), 224-239.

Ceccherini, S. (2005). Analytical determination of the regularization parameter in the retrieval of atmospheric vertical profiles. *Optics letters*, *30*(19), 2554-2556.

Ceccherini, S., Belotti, C., Carli, B., Raspollini, P., & Ridolfi, M. (2007). Regularization performances with the error consistency method in the case of retrieved atmospheric profiles. *Atmospheric Chemistry and Physics*, *7*(5), 1435-1440.

Chai, T., Carmichael, G. R., Tang, Y., Sandu, A., Heckel, A., Richter, A., & Burrows, J. P. (2009). Regional NOx emission inversion through a four-dimensional variational approach using SCIAMACHY tropospheric NO2 column observations. *Atmospheric Environment*, *43*(32), 5046-5055.

Chang, M. E., Hartley, D. E., Cardelino, C., & Chang, W. L. (1996). Inverse modeling of biogenic isoprene emissions. *Geophysical Research Letters*, *23*(21), 3007-3010.

Christiansen, B., Schmith, T., & Thejll, P. (2009). A surrogate ensemble study of climate reconstruction methods: Stochasticity and robustness. *Journal of Climate*, *22*(4), 951-976.

Chung, J., & Nagy, J. G. (2010). An efficient iterative approach for large-scale separable nonlinear inverse problems. *SIAM Journal on Scientific Computing*, *31*(6), 4654-4674.

Chung, J., Nagy, J. G., & O'Leary, D. P. (2008). A weighted GCV method for Lanczos hybrid regularization. *Electronic Transactions on Numerical Analysis*, *28*(149-167), 2008.

Cuccia, E., Bernardoni, V., Massabò, D., Prati, P., Valli, G., & Vecchi, R. (2010). An alternative way to determine the size distribution of airborne particulate matter. *Atmospheric Environment*, *44*(27), 3304-3313.

De Vito, S., Piga, M., Martinotto, L., & Di Francia, G. (2009). CO, NO2 and NOx

urban pollution monitoring with on-field calibrated electronic nose by automatic bayesian regularization. *Sensors and Actuators B: Chemical*, *143*(1), 182-191.

Doicu, A., Schreier, F., & Hess, M. (2004). Iterative regularization methods for atmospheric remote sensing. *Journal of Quantitative Spectroscopy and Radiative Transfer*, *83*(1), 47-61.

Doicu, A., Schreier, F., & Hess, M. (2005). An iterative regularization method with B-spline approximation for atmospheric temperature and concentration retrievals. *Environmental Modelling & Software*, *20*(9), 1101-1109.

Doicu, A., Schreier, F., Hilgers, S., & Hess, M. (2005). Multi-parameter regularization method for atmospheric remote sensing. *Computer physics communications*, *165*(1), 1-9.

Doicu, A., Schreier, F., Hilgers, S., & Hess, M. (2007). Error analysis and minimum bound method for atmospheric remote sensing. *Environmental Modelling & Software*, *22*(6), 837-846.

Doicu, A., Schüssler, O., & Loyola, D. (2010). Constrained regularization methods for ozone profile retrieval from UV/VIS nadir spectrometers. *Journal of Quantitative Spectroscopy and Radiative Transfer*, *111*(6), 907-916.

Drignei, D., Forest, C. E., & Nychka, D. (2008). Parameter estimation for computationally intensive nonlinear regression with an application to climate modeling. *The Annals of applied statistics*, 1217-1230.

Eckhardt, S., Prata, A. J., Seibert, P., Stebel, K., & Stohl, A. (2008). Estimation of the vertical profile of sulfur dioxide injection into the atmosphere by a volcanic eruption using satellite column measurements and inverse transport modeling. *Atmospheric Chemistry and Physics*, *8*(14), 3881-3897.

Ekstrom, M. P., & Rhoads, R. L. (1974). On the application of eigenvector expansions to numerical deconvolution. *Journal of Computational Physics*, *14*(4), 319-340.

Elbern, H., Strunk, A., Schmidt, H., & Talagrand, O. (2007). Emission rate and chemical state estimation by 4-dimensional variational inversion. *Atmospheric Chemistry and Physics*, *7*(14), 3749-3769.

inversion," Atmospheric Chemistry and Physics, vol. 7, pp. 3749-3769, 2007.

Englert, C. R., Schimpf, B., Birk, M., Schreier, F., Krocka, M., Nitsche, R. G., ... & Summers, M. E. (2000). The 2.5 THz heterodyne spectrometer THOMAS: Measurement of OH in the middle atmosphere and comparison with photochemical model results. *Journal of Geophysical Research: Atmospheres*, *105*(D17), 22211-22223.

Eriksson, P. (2000). Analysis and comparison of two linear regularization methods for passive atmospheric observations. *Journal of Geophysical Research: Atmospheres*, *105*(D14), 18157-18167.

Fan, S. M., Sarmiento, J. L., Gloor, M., & Pacala, S. W. (1999). On the use of regularization techniques in the inverse modeling of atmospheric carbon dioxide. *Journal of Geophysical Research: Atmospheres*, *104*(D17), 21503-21512.

Fedorova, A. A., Korablev, O. I., Bertaux, J. L., Rodin, A. V., Montmessin, F., Belyaev, D. A., & Reberac, A. (2009). Solar infrared occultation observations by SPICAM experiment on Mars-Express: Simultaneous measurements of the vertical distributions of H2O, CO2 and aerosol. *Icarus*, *200*(1), 96-117.

Fierro, R. D., Golub, G. H., Hansen, P. C., & O'Leary, D. P. (1997). Regularization by truncated total least squares. *SIAM Journal on Scientific Computing*, *18*(4), 1223-1241.

Friedrich, V., & Hofmann, B. (1987). A predictor-corrector technique for constrained least-squares regularization. *Numerische Mathematik*, *51*(3), 353-367.

Frommer, A., & Maass, P. (1999). Fast CG-based methods for Tikhonov--Phillips regularization. *SIAM Journal on Scientific Computing*, *20*(5), 1831-1850.

Gaikovich, K. P. (1994, August). Tikhonov's method of the ground-based radiometric retrieval of the ozone profile. In *Proceedings of IGARSS'94-1994 IEEE International Geoscience and Remote Sensing Symposium* (Vol. 4, pp. 1901-1903). IEEE.

Gilliland, A. B., Appel, K. W., Pinder, R. W., & Dennis, R. L. (2006). Seasonal NH3 emissions for the continental united states: Inverse model estimation and evaluation. *Atmospheric Environment*, *40*(26), 4986-4998.

Golub, G. H., Heath, M., & Wahba, G. (1979). Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, *21*(2), 215-223.

Gustafsson, J., & Protas, B. (2010). Regularization of the backward-in-time Kuramoto–Sivashinsky equation. *Journal of computational and applied mathematics*, *234*(2), 398-406.

Haber, E., & Oldenburg, D. (2000). A GCV based method for nonlinear ill-posed problems. *Computational Geosciences*, *4*(1), 41-63.

Hansen, P. C. (1990). The discrete Picard condition for discrete ill-posed problems. *BIT Numerical Mathematics*, *30*(4), 658-672.

Hansen, P. C. (1994). Regularization tools: A Matlab package for analysis and solution of discrete ill-posed problems. *Numerical algorithms*, *6*(1), 1-35.

Hansen, P. C. (1998). *Rank-deficient and discrete ill-posed problems: numerical aspects of linear inversion*. Society for Industrial and Applied Mathematics.

Hansen, P. C., Jensen, T. K., & Rodriguez, G. (2007). An adaptive pruning algorithm for the discrete L-curve criterion. *Journal of computational and applied mathematics*, *198*(2), 483-492.

Hansen, P. C., & O'Leary, D. P. (1993). The use of the L-curve in the regularization of discrete ill-posed problems. *SIAM journal on scientific computing*, *14*(6), 1487-1503.

Hasekamp, O. P., & Landgraf, J. (2001). Ozone profile retrieval from backscattered ultraviolet radiances: The inverse problem solved by regularization. *Journal of Geophysical Research: Atmospheres*, *106*(D8), 8077-8088.

Henze, D. K., Seinfeld, J. H., & Shindell, D. T. (2009). Inverse modeling and mapping US air quality influences of inorganic PM 2.5 precursor emissions using the adjoint of GEOS-Chem. *Atmospheric Chemistry and Physics*, *9*(16), 5877-5903.

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for

nonorthogonal problems. *Technometrics*, *12*(1), 55-67.

Hoerl, A. E., & Kennard, R. W. (1976). Ridge regression iterative estimation of the biasing parameter. *Communications in Statistics-Theory and Methods*, *5*(1), 77-88.

Horenko, I. (2010). On clustering of non-stationary meteorological time series. *Dynamics of Atmospheres and Oceans*, *49*(2-3), 164-187.

Jensen, T. K., & Hansen, P. C. (2007). Iterative regularization with minimum-residual methods. *BIT Numerical Mathematics*, *47*(1), 103-120.

Kalnay, E., Li, H., Miyoshi, T., Yang, S. C., & Ballabrera-Poy, J. (2007). 4-D-Var or ensemble Kalman filter?. *Tellus A: Dynamic Meteorology and Oceanography*, *59*(5), 758-773.

Kaufman, L., & Neumaier, A. (1997). Regularization of ill-posed problems by envelope guided conjugate gradients. *Journal of Computational and Graphical Statistics*, *6*(4), 451-463.

Kilmer, M. E., & O'Leary, D. P. (2001). Choosing regularization parameters in iterative methods for ill-posed problems. *SIAM Journal on matrix analysis and applications*, *22*(4), 1204-1221.

Kitagawa, T., Nakata, S., & Hosoda, Y. (2001). Regularization using QR factorization and the estimation of the optimal parameter. *BIT Numerical Mathematics*, *41*(5), 1049-1058.

Koner, P. K., & Drummond, J. R. (2008). A comparison of regularization techniques for atmospheric trace gases retrievals. *Journal of Quantitative Spectroscopy and Radiative Transfer*, *109*(3), 514-526.

Koner, P. K., & Drummond, J. R. (2008). Atmospheric trace gases profile retrievals using the nonlinear regularized total least squares method. *Journal of Quantitative Spectroscopy and Radiative Transfer*, *109*(11), 2045-2059.

Krakauer, N. Y., Schneider, T., Randerson, J. T., & Olsen, S. C. (2004). Using generalized cross-validation to select parameters in inversions for regional carbon fluxes. *Geophysical research letters*, *31*(19).

Krawczyk-StanDo, D., & Rudnicki, M. (2007). Regularization parameter selection in discrete ill-posed problems--The use of the U-Curve. *International Journal of*

Applied Mathematics and Computer Science, *17*(2), 157.

Landi, G. (2008). The Lagrange method for the regularization of discrete ill-posed problems. *Computational Optimization and Applications*, *39*(3), 347-368.

Lau, K. T., Guo, W., Kiernan, B., Slater, C., & Diamond, D. (2009). Non-linear carbon dioxide determination using infrared gas sensors and neural networks with Bayesian regularization. *Sensors and Actuators B: Chemical*, *136*(1), 242-247.

Lemmetty, M., Keskinen, J., & Marjamäki, M. (2005). The ELPI response and data reduction II: Properties of kernels and data inversion. *Aerosol science and technology*, *39*(7), 583-595.

Lewis, J. M., Lakshmivarahan, S., & Dhall, S. (2006). *Dynamic data assimilation: a least squares approach* (Vol. 13). Cambridge University Press.

Li, M. J., Chen, D. S., Cheng, S. Y., Wang, F., Li, Y., Zhou, Y., & Lang, J. L. (2010). Optimizing emission inventory for chemical transport models by using genetic algorithm. *Atmospheric Environment*, *44*(32), 3926-3934.

Lin, Y., & Wohlberg, B. (2008, March). Application of the UPRE method to optimal parameter selection for large scale regularization problems. In *2008 IEEE Southwest Symposium on Image Analysis and Interpretation* (pp. 89-92). IEEE.

Lorenz, D. A., & Rösch, A. (2010). Error estimates for joint Tikhonov and Lavrentiev regularization of constrained control problems. *Applicable Analysis*, *89*(11), 1679-1691.

Mead, J. L. (2008). Parameter estimation: A new approach to weighting a priori information. *J. Inv. Ill-posed Problems*, *16*(2), 175-194.

Meijer, Y. J., Swart, D. P. J., Baier, F., Bhartia, P. K., Bodeker, G. E., Casadio, S., ... & Zehner, C. (2006). Evaluation of Global Ozone Monitoring Experiment (GOME) ozone profiles from nine different algorithms. *Journal of Geophysical Research: Atmospheres*, *111*(D21).

Mendoza-Dominguez, A., & Russell, A. G. (2000). Iterative inverse modeling and direct sensitivity analysis of a photochemical air quality model. *Environmental science & technology*, *34*(23), 4974-4981.

Mendoza-Dominguez, A., & Russell, A. G. (2001). Emission strength validation using four-dimensional data assimilation: Application to primary aerosol and precursors to ozone and secondary aerosol. *Journal of the Air & Waste Management Association*, *51*(11), 1538-1550.

Mendoza-Dominguez, A., & Russell, A. G. (2001). Estimation of emission adjustments from the application of four-dimensional data assimilation to photochemical air quality modeling. *Atmospheric Environment*, *35*(16), 2879-2894.

Menke, W. (1989). *Geophysical data analysis: Discrete inverse theory*. Academic press.

Morozov, V. A. (1966). On the solution of functional equations by the method of regularization. In *Doklady Akademii Nauk* (Vol. 167, No. 3, pp. 510-512). Russian Academy of Sciences.

Napelenok, S. L., Pinder, R. W., Gilliland, A. B., & Martin, R. V. (2008). A method for evaluating spatially-resolved NO x emissions using Kalman filter inversion, direct sensitivities, and space-based NO 2 observations. *Atmospheric Chemistry and Physics*, *8*(18), 5603-5614.

Neumaier, A. (1998). Solving ill-conditioned and singular linear systems: A tutorial on regularization. *SIAM review*, *40*(3), 636-666.

Nikoukar, R., Kamalabadi, F., Kudeki, E., & Sulzer, M. (2008). An efficient near-optimal approach to incoherent scatter radar parameter estimation. *Radio Science*, *43*(05), 1-15.

Osterloh, L., Pérez, C., Böhme, D., Baldasano, J. M., Böckmann, C., Schneidenbach, L., & Vicente, D. (2009). Parallel software for retrieval of aerosol distribution from LIDAR data in the framework of EARLINET-ASOS. *Computer Physics Communications*, *180*(11), 2095-2102.

Pétron, G., Granier, C., Khattatov, B., Lamarque, J. F., Yudin, V., Müller, J. F., & Gille, J. (2002). Inverse modeling of carbon monoxide surface emissions using Climate Monitoring and Diagnostics Laboratory network observations. *Journal of Geophysical Research: Atmospheres*, *107*(D24), ACH-10.

Qin, Y., Box, M. A., & Jupp, D. L. (2002). Inversion of multiangle sky radiance measurements for the retrieval of atmospheric optical properties 1. Algorithm. *Journal of Geophysical Research: Atmospheres*, *107*(D22), AAC-10.

Quélo, D., Mallet, V., & Sportisse, B. (2005). Inverse modeling of NOx emissions at regional scale over northern France: Preliminary investigation of the second-order sensitivity. *Journal of Geophysical Research: Atmospheres*, *110*(D24).

Rezghi, M., & Hosseini, S. M. (2009). A new variant of L-curve for Tikhonov regularization. *Journal of computational and applied mathematics*, *231*(2), 914-924.

Riefler, N., & Wriedt, T. (2008). Intercomparison of Inversion Algorithms for Particle-Sizing Using Mie Scattering. *Particle & Particle Systems Characterization*, *25*(3), 216-230.

Rozanov, A., Eichmann, K. U., Von Savigny, C., Bovensmann, H., Burrows, J. P., Von Bargen, A., ... & McDermid, I. S. (2007). Comparison of the inversion algorithms applied to the ozone vertical profile retrieval from SCIAMACHY limb measurements. *Atmospheric Chemistry and Physics*, *7*(18), 4763-4779.

Rust, B. W. (1998). *Truncating the singular value decomposition for ill-posed problems*. US Department of Commerce, Technology Administration, National Institute of Standards and Technology.

Rust, B. W. (2000). Parameter selection for constrained solutions to ill-posed problems. *Computing science and statistics*, *32*, 333-347.

Rust, B. W., & O'Leary, D. P. (2008). Residual periodograms for choosing regularization parameters for ill-posed problems. *Inverse Problems*, *24*(3), 034005.

Saad, Y., & Schultz, M. H. (1986). GMRES: A generalized minimal residual algorithm for solving nonsymmetric linear systems. *SIAM Journal on scientific and statistical computing*, *7*(3), 856-869.

Saide, P., Osses, A., Gallardo, L., & Osses, M. (2009). Adjoint inverse modeling of a CO emission inventory at the city scale: Santiago de Chile's

case. *Atmospheric Chemistry and Physics Discussions*, *9*(2), 6325-6361.

Salcedo-Sanz, S., Portilla-Figueras, J. A., Ortiz-Garcia, E. G., Perez-Bellido, A. M., Garcia-Herrera, R., & Elorrieta, J. I. (2009). Spatial regression analysis of NOx and O3 concentrations in Madrid urban area using Radial Basis Function networks. *Chemometrics and Intelligent Laboratory Systems*, *99*(1), 79-90.

Santamarina, J. C., & Fratta, D. (1998). *Introduction to discrete signals and inverse problems in civil engineering*.

Schimpf, B., & Schreier, F. (1997). Robust and efficient inversion of vertical sounding atmospheric high-resolution spectra by means of regularization. *Journal of Geophysical Research: Atmospheres*, *102*(D13), 16037-16055.

Smith, C. R., & Grandy Jr, W. T. (Eds.). (2013). *Maximum-Entropy and bayesian methods in inverse problems* (Vol. 14). Springer Science & Business Media.

Stark, P. B., & Parker, R. L. (1995). Bounded-variable least-squares: an algorithm and applications. *Computational Statistics*, *10*, 129-129.

Stohl, A., Seibert, P., Arduini, J., Eckhardt, S., Fraser, P., Greally, B. R., ... & Yokouchi, Y. (2009). An analytical inversion method for determining regional and global emissions of greenhouse gases: Sensitivity studies and application to halocarbons. *Atmospheric Chemistry and Physics*, *9*(5), 1597-1620.

Talukdar, S. S., & Swihart, M. T. (2003). An improved data inversion program for obtaining aerosol size distributions from scanning differential mobility analyzer data. *Aerosol Science and Technology*, *37*(2), 145-161.

Vogel, C. R. (2002). *Computational methods for inverse problems*. Society for Industrial and Applied Mathematics.

Voutilainen, A. (2001). *Statistical inversion methods for the reconstruction of aerosol size distributions*. Finnish Association for Aerosol Research.

Voutilainen, A., Stratmann, F., & Kaipio, J. P. (2000). A non-homogeneous regularization method for the estimation of narrow aerosol size distributions. *Journal of aerosol science*, *31*(12), 1433-1445.

Wang, Y. (2008). An efficient gradient method for maximum entropy regularizing retrieval of atmospheric aerosol particle size distribution function. *Journal of Aerosol Science*, *39*(4), 305-322.

Wang, Y., Fan, S., & Feng, X. (2007). Retrieval of the aerosol particle size distribution function by incorporating a priori information. *Journal of Aerosol Science*, *38*(8), 885-901.

Wang, Y., Fan, S., Feng, X., Yan, G., & Guan, Y. (2006). Regularized inversion method for retrieval of aerosol particle size distribution function in W 1, 2 space. *Applied optics*, *45*(28), 7456-7467.

Wang, Y., Liang, G., & Pan, Z. (2010). Inversion of particle size distribution from light-scattering data using a modified regularization algorithm. *Particuology*, *8*(4), 365-371.

Willoughby, R. A. (1979). Solutions of ill-posed problems (AN Tikhonov and VY Arsenin). *SIAM Review*, *21*(2), 266.

Wolfenbarger, J. K., & Seinfeld, J. H. (1990). Inversion of aerosol size distribution data. *Journal of Aerosol Science*, *21*(2), 227-247.

Wu, L. (2003). A parameter choice method for Tikhonov regularization. *Electronic Transactions on Numerical Analysis*, *16*, 107-128.

Xiao, X., Prinn, R. G., Fraser, P. J., Weiss, R. F., Simmonds, P. G., O'Doherty, S., ... & Cunnold, D. M. (2010). Atmospheric three-dimensional inverse modeling of regional industrial emissions and global oceanic uptake of carbon tetrachloride. *Atmospheric Chemistry and Physics*, *10*(21), 10421-10434.

Zhao, H., Xu, W., Yang, H., Li, X., Wang, J., & Cui, H. (2003, July). The maximum entropy algorithm for the determination of the Tikhonov regularization parameter in quantitative remote sensing inversion. In *IGARSS 2003. 2003 IEEE International Geoscience and Remote Sensing Symposium. Proceedings (IEEE Cat. No. 03CH37477)* (Vol. 6, pp. 3875-3877). IEEE.

This page intentionally left blank

Chapter 4

Application of direct regularization techniques and bounded-variable least squares for inverse modeling of an urban emissions inventory

This chapter was published in its current form, as a research article by Ana Yael Vanoye García and Alberto Mendoza Domínguez, in journal Atmospheric Pollution Research, Volume 5, Issue 2, April 2014, pages 219-225, doi.org/10.5094/APR.2014.027. The regularization techniques used in this article were selected based on the information presented in Chapter 3 of this dissertation.

4.1 Abstract

Inverse modeling, coupled with comprehensive air quality models, is being increasingly used for improving spatially and temporally resolved emissions inventories. Of the techniques available to solve the corresponding inverse problem, regularization techniques can provide stable solutions. However, in many instances, it is not clear which regularization parameter selection method should be used in conjunction with a particular regularization technique to get the best results. In this work, three regularization techniques (Tikhonov regularization, truncated singular-value decomposition, and damped singular-value decomposition) and three regularization parameter selection methods (generalized cross validation, the L-curve method [LC], and normalized cumulative periodograms) were applied in conjunction with an air quality model with the aim of identifying the best combination of regularization technique and parameter selection method when using inverse modeling to identify possible flaws in an urban-scale emissions inventory. The bounded-variable least-squares method (BVLS), which is not usually considered a regularization method, was also investigated. The results indicate that the choice of the regularization parameter explains most of the differences between the regularization techniques used, with the LC method exhibiting the best performance for the application described here.

The results also show that the BVLS scheme provides the best agreement between the observed and modeled concentrations among the mathematical techniques tested.

Keywords: air quality model, photochemical modeling, emissions evaluation, inverse problem

4.2 Introduction

Three-dimensional comprehensive air quality models (AQMs) describe the atmospheric transport and transformation of trace species and are routinely used in the development of pollution-reduction strategies and other air quality management policies. However, to produce accurate, trustworthy results, AQMs rely on the use of detailed emission inventories that, even today, convey a great degree of uncertainty (Miller et al., 2006; Russell, 2008). This translates into model applications in which discrepancies between model-derived concentrations and observations of air pollutants can be quite large. In this description, we are assuming that the AQM is perfect, and the emissions are one of the most uncertain input parameters in the modeling effort (e.g. Russell and Dennis, 2000; Tian et al., 2010).

One way of reducing emission inventory uncertainty is to use inverse modeling or data assimilation techniques to identify possible flaws in the construction of such emission inventories. Applications of inverse modeling range from global (e.g., Pétron et al., 2002) or continental (e.g., Elbern et al., 2007) to regional or urban scales (e.g., Quélo et al., 2005) for a variety of atmospheric species such as stratospheric ozone depletion substances (e.g., Xiao et al., 2010), greenhouse gases (e.g., Stohl et al., 2009), radioactive material (e.g., Winiarek et al., 2012), and tropospheric ozone and aerosol precursors (e.g., Gilliland et al., 2006; Napelenok et al., 2008; Henze et al., 2009). In this context, several mathematical techniques have been used to find solutions for the corresponding inverse

problems, including four-dimensional data assimilation (e.g., Meirink et al., 2008), Kalman or ensemble Kalman filtering (e.g., Napelenok et al., 2008), and the use of adjoint models (e.g., Hakami et al., 2005). Li et al. (2010) used genetic algorithms for optimizing inventories, but their application was limited because of the necessary computational requirements.

One approach for performing this top-down emissions inventory evaluation is to first use a forward model (the AQM) to compute both the simulated concentration fields of pollutants and their responses to changes in emissions (sensitivity fields). With this, a linear model of the form $\mathbf{Gm} = \mathbf{d}$ can be constructed, where $\mathbf{d}$ is a vector containing the difference between modeled and observed concentrations, $\mathbf{G}$ is a matrix containing the sensitivity coefficients of all pollutant species to changes in the emission strengths, and $\mathbf{m}$ is a vector of emission strength changes that brings the observations and model-derived concentrations into agreement. Then, if an over-determined least-squares problem is solved, the corresponding inverse model can be represented as $\mathbf{m}^{est} = (\mathbf{G}^{\mathrm{T}}\mathbf{G})^{-1}\mathbf{G}^{\mathrm{T}}\mathbf{d}$, where $\mathbf{G}^{\mathrm{T}}$ is the transpose of $\mathbf{G}$. However, inverse problems are typically ill conditioned, and this is an inconvenience because in practice, observations often possess a certain degree of error or noise (Aster et al., 2005).

Several mathematical techniques based on the incorporation of known properties about the solution have been developed with the aim of improving the conditioning of direct inverse problems, including regularization. However, there are few examples of the use of regularization techniques to obtain inverse-derived emissions inventories. In particular, few formal methods have been applied to obtain the value of the regularization parameters. This paper addresses the issue of performing inverse modeling of an urban air-pollutant emission inventory by applying direct regularization techniques. Three regularization techniques and three regularization parameter choice methods were assessed. An additional technique investigated, which is not usually considered a regularization method, was the bounded-variable least-squares (BVLS) method.

4.3 Methods

4.3.1 Regularization methods

Three regularization methods were explored in this work: Tikhonov regularization (TIKH), truncated singular-value decomposition (TSVD), and damped singular-value decomposition (DSVD). Tikhonov's method consists of substituting the least-squares problem for a problem of the form (Neumaier, 1998):

$$\mathbf{m}^{est} = \min\left\{\left\|\mathbf{Gm}-\mathbf{d}\right\|_2 + \left\|\mathbf{Lm}\right\|_2\right\} \tag{4.1}$$

where $\mathbf{L} = \lambda\mathbf{I}$, $\mathbf{I}$ is typically the identity matrix, and $\lambda \in \mathbb{R}$ is a regularization parameter that controls the weight given to the minimization of the additional restriction relative to the minimization of the residual norm. Thus, TIKH seeks a solution that minimizes a criterion made up of the sum of two components: a weighted least-square term and a quadratic penalty term on the solution.

The singular-value decomposition of matrix **G** with r = rank(**G**), as in the following equation:

$$\mathbf{G} = \mathbf{U}_r\mathbf{S}_r\mathbf{V}_r^T = \sum_{i=1}^{r}\boldsymbol{\sigma}_i\mathbf{u}_i\mathbf{v}_i^T \tag{4.2}$$

can be used to obtain the Moore-Penrose pseudo-inverse of $\mathbf{G}$. However, the generalized inverse solution can become unstable when some of the singular values, $\boldsymbol{\sigma}_i$, are small (Aster et al., 2005). Therefore, in TSVD, it is assumed that it is possible to recover a useful model by truncating the sum in Equation (4.2) in an upper-bound $k < r$ before the smallest singular values start dominating (Hansen, 1990). When $k = r$, the solution of TSVD is identical to the solution obtained by ordinary least-square methods. However, a solution obtained from TSVD with $k < r$ will tend to be more stable (Aster et al., 2005). Finally, DSVD (Ekstrom and Rhodes, 1974) may be regarded as a regularization method that follows Tikhonov in terms of its TSVD, with the difference being that DSVD introduces a smoother

cutoff by means of filter factors that decay slower than the Tikhonov method, overall requiring less filtering (Hansen, 1998; Lin et al., 2011).

4.3.2 Regularization parameter choice methods

Although the proper choice of the regularization parameter (either the continuous parameter λ or the discrete parameter k) is essential for the effectiveness of the regularization methods applied (Hansen, 1998), the optimal determination of this parameter remains an open issue (Krawczyk-Stando and Rudnicki, 2007; Lin et al., 2011). Some previous studies where inverse modeling has been applied to evaluate emissions inventories have used formal methods to obtain the value of the regularization parameter (Fan et al., 1999; Mendoza-Dominguez and Russell, 2001; Krakauer et al., 2004; Chai et al., 2009; Henze et al., 2009; Saide et al., 2009). However, the strategy of assigning values to the regularization parameter subjectively, or empirically, prevails (e.g., Eckhardt et al., 2008).

In this study, we explore three methods that do not require a good estimate of the noise variance: generalized cross validation (GCV) (Golub et al., 1979; Haber and Olenburg, 1999), which is a parameter-choice method based on ordinary cross validation (Allen, 1974); the L-curve method (LC), which uses a plot of the valid regularization parameters of the (semi) norm of the regularized solution versus the corresponding residual norm (the best regularization parameter must be located in the corner of the L-curve) (Hansen and O'Leary, 1993); and normalized cumulative periodograms (NCP) (Rust, 2000; Rust and O'Leary, 2008), which chooses the regularization parameter for which the residual becomes closer to behave as white Gaussian noise.

4.3.3 Regularization with restrictions

Regardless of whether regularization methods tend to yield more stable, precise solutions, these solutions will often lack physical sense or violate some of the restrictions imposed by the nature of the problem. In our case we need to guarantee positive emissions. Several techniques incorporate additional

restrictions to impose boundaries for the solution or add additional information about the solution. One such technique is the BVLS (Stark and Parker, 1995), which solves linear least-squares problems with upper and lower bounds on the variables.

BVLS uses an active set strategy in which the unconstrained least-squares problems for each candidate set of free variables are solved using the QR decomposition. The method also includes a “warm-start” feature that accelerates the solution by allowing for some of the variables to initialize at their upper or lower bounds. Stark and Parker’s BVLS algorithm is based on the non-negative least squares method (Lawson and Hanson, 1974). In this study, we use this additional technique in our application and compare it with the solutions obtained by regularization.

4.4 Application to the GMA emissions inventory

The base case, reported by Mendoza and García (2009), for the modeling of photochemical pollutants in the Guadalajara Metropolitan Area (GMA; 20° 40’ 25” N, 103° 20’ 38” W) was used as case study. The GMA is the second largest urban center in Mexico, emissions are rather concentrated around the urban core (~600 km^2), and no important emission sources are located around this core. In the application described here, the same AQM, spatial configuration of the modeling domain, as well as the same meteorological fields, emissions inventory, and initial and boundary conditions were used. In that study, the California/Carnegie Institute of Technology (CIT) model extended with the capacity of estimating first-order sensitivity coefficients through the direct-decoupled method for three-dimensional models (Yang et al., 1997) was applied for the simulation of a three-day high-ozone concentration episode, occurring from May 16 to May 18, 2001. May was selected for the modeling exercise because it is the month when O_3 concentrations peak in the GMA (Zuk et al., 2007). The modeling domain was a computational matrix composed of 40 × 40 cells (horizontal resolution), with each cell being 4 × 4 km

and geographically centered in the GMA. In addition, the domain included six vertical levels topping at 3100 m.

The emission inventory used by Mendoza and Garcia (2009) was based on the 1995 Official Emissions Inventory for the GMA. This inventory had to be extended to provide coverage for the additional municipalities that were included in the modeling domain and that were not part of the GMA. In addition, this inventory had to be scaled from the base year (1995) to the modeled year (2001) and had to be spatially segregated based on an estimated population density. For this research, it was particularly suitable to have an emission inventory that, given its formulation process, was known to have uncertainties (Mendoza and Garcia, 2011).

Previously, Mendoza and Garcia (2011) used ridge regression (Hoerl and Kennard, 1976) to derive hour-to-hour inverse-modeled emission scaling factors to the emissions for this same application. They found that on a daily average, CO emissions would need to be subject to corrections ranging from −16% to +60%, whereas NO_x and SO_2 emissions would require increments from 100% to 150% and 20% to 140%, respectively. The inverse model proposed by Mendoza and García (2011) notably enhanced the statistical model performance for O_3 and other pollutants predictions; however, several discrepancies among the observed and modeled values remained unsolved.

This new application recreates the original problem, as approached by Mendoza and García (2011), but differs from it in the mathematical methods used. Eight different schemes were tested and compared with the base case and inverse-derived results achieved by Mendoza and García (2011). The schemes were as follows: TIKH combined with GCV, LC, and NCP; DSVD combined with GCV, LC, and NCP; TSVD combined with GCV; and BVLS. The algorithms provided in the Regularization Tools Matlab package, developed by Hansen (2008), were used to perform our numerical experiments. For the BVLS method, a modified version of the original FORTRAN algorithm by Stark and Parker (1995) was used.

The inversion experiments consisted of obtaining correction factors for domain-wide emission of NO_x, CO, and SO_2 using the observational data derived by the eight ground monitoring stations that comprise the routine air quality network of the GMA. Observations of NO, NO_2, CO, SO_2, and O_3 were used in this process. Because the inverse model works under a minimization scheme difference between the observed and simulated values, leaving a residual error that the model cannot explain, and because of the structure of the minimization function, biased estimators are obtained. Therefore, a complete concordance between the observed and modeled values after the inversion process is completed cannot be anticipated. However, a better performance of the AQM may be expected not only for the species directly related to the emissions (i.e., NO, NO_2, CO, and SO_2), but also for the secondary species (e.g., O_3). The inverse-modeling approach, as used here, yielded hourly changes to the original emission inventory that needed to be applied for the AQM to more successfully replicate the observed atmospheric pollutant concentrations. The inversion was conducted under an iterative process due to the non-linear response of some of the constituents (particularly NO, NO_2, and O_3) to the changes in emissions.

Finally, when conducting inverse modeling of emissions, it must be considered that misfits between the model and observations are due to not only emission inaccuracies, but also to errors in meteorological fields and other model parameters, as well as errors in the representation of physical and chemical processes. For this reason, a model evaluation is often required before the inverse-modeling stage (Saide et al., 2009) or these errors must be accounted for by incorporating them into the methodology as an additional term to the minimization function (e.g. Elbern et al., 2007). The corresponding evaluation processes performed for the meteorological and AQMs used in this study were previously described by Mendoza and García (2009). Both models were found to perform within the recommended guidelines for related applications.

4.5 Results

Table 4.1 depicts the statistical performance of the base case, that is, the ability of the CIT model to replicate O_3-ambient concentrations. For brevity, only third-day (May 18) simulation results are shown. Table 4.1 shows the contrasts of that performance with the Mendoza and García (2011) experiment (referred to as the MG test for the remainder of the text) and the eight schemes investigated. The statistical indicators for performance appraisal were those suggested by Doll et al. (1991) for use in AQM applications. Moreover, Tables 4.2, 4.3, and 4.4 show the statistical performance of the AQM for the same day for each of the chemical species whose emissions where directly adjusted by the inverse-modeling process, namely, NO_x, CO, and SO_2, using the previously mentioned inverse-modeling schemes.

As shown in Tables 4.1 through 4.4, among the regularization and restricted least squares methods tested, for most performance metrics and for all the chemical species of interest, BVLS consistently performed the best. Particularly, for O_3, Doll et al. (1991) suggest the following performance benchmarks: a normalized bias smaller than ±15%, normalized error smaller than ±35%, and peak estimation accuracy smaller than ±20%. All models tested within these limits, except for the normalized bias, for which only BVLS, DSVD-NCP, and TIKH-NCP yielded satisfactory results. However, BVLS performance did not always outdo that reported in the MG case, especially concerning O_3 concentrations, but it improved SO_2 significantly (e.g., the daily index of agreement [DIA] increased from 0.48 to 0.52). CO and NO_x simulations were also enhanced except for the normalized bias, mean normalized square error (MNSE), and root mean square error (RMSE) of NO_x. MNSE is usually regarded as a better metric for spatial and temporal performance appraisal than the normalized error (Hanna, 1998).

One of the main differences between regularization methods and BVLS is that the latter restricts the solution to a definite, fixed set of maximum and minimum values,

while regularization methods seek a smoother solution. It is possible that BVLS's relatively good performance might be a result of the inventory attributes and that the need of imposing appropriate, plausible bounds was of major concern. This possibility could be explored by performing further experiments with other inventories.

All methods generally showed lower RMSE when compared with the base case. However, it is important to also assess the relative weights of the systematic ($RMSE_s$) and unsystematic ($RMSE_u$) components of the RMSE. In this context, the regularization schemes tested yielded mixed results. In all NO_x and SO_2 simulations, $RMSE_s$ prevailed over $RMSE_u$, whereas for the CO results, $RMSE_u$ outweighed its systematic counterpart. O_3 showed a mixed response: only BVLS and those simulations using LC and NCP parameter choice schemes presented higher $RMSE_u$ than $RMSE_s$, indicating that in such applications, residual errors are mostly caused by variations (noise) that the forward model cannot resolve.

Time series for the RMSE statistic were also explored with the aim of determining whether the CIT was adequately simulating the temporal dynamics of O_3. For the BVLS run, Figure 4.1 shows the time evolution of RMSE, $RMSE_s$, and $RMSE_u$. It should be noted that overall, RMSE diminished at all hours compared with the base case simulation. Furthermore, as was also reported in the MG case, there appears to be a local minimum in the RMSE during the afternoon hours. This might be a result of the fact that the CIT, as other photochemical air quality models, has been partially calibrated (tuned) to perform best under conditions were O_3 levels are the highest as O_3 peaks are of more environmental concern than low concentrations (Russell and Dennis, 2000). However, the mere presence of $RMSE_s$ indicates that the model requires further improvement for better results.

Table 4.1. CIT statistical performance evaluation for simulated O_3 on May 18, 2001[a]

	BC	MG	BVLS	DSVD-GCV	DSVD-LC	DSVD-NCP	TIKH-GCV	TIKH-LC	TIKH-NCP	TSVD-GCV
Peak estimation accuracy, %	16.40	–10.20	–6.24	–6.80	–1.66	–4.05	2.23	–1.46	–6.96	3.15
Normalized bias, %	20.30	0.10	11.14	16.61	16.96	13.50	23.54	16.27	7.08	23.79
Normalized error	36.50	15.60	21.41	31.81	26.94	23.01	35.27	26.06	18.81	35.56
MNSE	0.19	0.04	0.05	0.13	0.08	0.06	0.16	0.08	0.04	0.16
RMSE (ppbv)	31.10	21.90	22.65	29.47	24.41	23.36	32.16	24.29	22.95	32.85
$RMSE_s$ (ppbv)	20.70	15.10	14.24	20.98	15.39	15.28	23.55	16.09	16.11	24.56
$RMSE_u$ (ppbv)	23.20	15.80	17.59	20.70	18.94	17.66	21.90	18.20	16.35	21.81
DIA	0.89	0.95	0.95	0.90	0.94	0.94	0.88	0.94	0.94	0.87

[a] Statistics were computed by taking into account the residual $r_i = P_i - O_i$, where O_i and P_i are the i-th observed and modeled concentrations, respectively. Normalized bias is $1/N\Sigma(r_i/O_i)$, where N represents the number of valid pairs that originate r_i while the sum runs from i=1 to N. In a similar fashion, the normalized error is $1/N\Sigma[|r_i|/O_i]$, the *MNSE* is $1/N\Sigma(r_i/O_i)^2$, and the *RMSE* is $[1/N\Sigma(r_i)^2]^{1/2}$. $RMSE_s$ was computed from $[1/N\Sigma(\underline{r}_i)^2]^{1/2}$, where $\underline{r}_i = \underline{P}_i - O_i$ and $\underline{P}_i = a + bO_i$ (a and b are lineal regression coefficients). *RMSE* follows $RMSE^2 = RMSE_s^2 + RMSE_u^2$. Finally, the *DIA* is $1 - [N\,RMSE^2/\Sigma(|P_i - M_o| + |O_i - M_o|)^2]$, where M_o is the mean observed value, as given by $1/N\Sigma O_i$. Note that a positive bias indicates that modeled values are greater than observed values.

Table 4.2. CIT statistical performance evaluation for simulated NO_x on May 18, 2001[a]

	BC	MG	BVLS	DSVD-GCV	DSVD-LC	DSVD-NCP	TIKH-GCV	TIKH-LC	TIKH-NCP	TSVD-GCV
Normalized bias, %	–51.10	–13.20	–29.41	–56.84	–32.66	–33.35	–53.77	–35.33	–34.07	–55.90
Normalized error	67.00	61.90	–20.38	69.29	66.04	64.72	68.44	63.26	62.23	69.55
MNSE	2.05	0.83	1.05	2.45	1.15	1.13	2.38	1.18	1.11	2.56
RMSE (ppbv)	43.10	36.70	37.94	44.36	38.65	38.42	44.56	38.46	38.14	45.06
$RMSE_s$ (ppbv)	39.70	26.10	28.14	41.31	29.93	29.23	41.63	30.79	28.69	42.39
$RMSE_u$ (ppbv)	16.90	25.80	25.45	16.16	24.47	24.94	15.89	23.05	25.12	15.30
DIA	0.47	0.59	0.60	0.46	0.58	0.59	0.45	0.58	0.60	0.44

[a] Refer to Table 4.1 for definitions.

Table 4.3. CIT statistical performance evaluation for simulated CO on May 18, 2001[a]

	BC	MG	BVLS	DSVD-GCV	DSVD-LC	DSVD-NCP	TIKH-GCV	TIKH-LC	TIKH-NCP	TSVD-GCV
Normalized bias, %	–26.80	–15.60	–15.12	–20.96	–14.88	–15.04	–15.53	–15.21	–15.23	–15.34
Normalized error	60.60	46.20	45.05	45.12	45.29	45.28	45.32	45.31	45.32	45.27
MNSE	0.87	0.44	0.44	0.44	0.44	0.44	0.44	0.44	0.44	0.44
RMSE (ppmv)	1.33	1.05	1.04	1.04	1.04	1.04	1.04	1.04	1.04	1.04
$RMSE_s$ (ppmv)	1.01	0.62	0.64	0.64	0.64	0.64	0.65	0.64	0.64	0.65
$RMSE_u$ (ppmv)	0.86	0.84	0.82	0.82	0.82	0.82	0.82	0.82	0.82	0.82
DIA	0.60	0.76	0.76	0.76	0.76	0.76	0.76	0.76	0.76	0.76

[a] Refer to Table 4.1 for definitions.

Table 4.4. CIT statistical performance evaluation for simulated SO_2 on May 18, 2001[a]

	BC	MG	BVLS	DSVD-GCV	DSVD-LC	DSVD-NCP	TIKH-GCV	TIKH-LC	TIKH-NCP	TSVD-GCV
Normalized bias, %	–13.10	10.10	23.85	–16.67	4.41	14.76	–20.51	–10.20	–4.22	–20.25
Normalized error	66.90	79.90	82.35	64.95	69.83	77.22	63.82	64.76	69.60	64.07
MNSE	1.99	1.43	1.11	2.10	1.35	1.26	2.25	1.81	1.70	2.25
RMSE (ppbv)	8.10	7.70	7.26	8.13	7.34	7.44	8.23	7.85	7.89	8.25
$RMSE_s$ (ppbv)	7.30	6.40	5.98	7.43	6.51	6.36	7.58	7.18	7.01	7.57
$RMSE_u$ (ppbv)	3.50	4.20	4.11	3.30	3.37	3.85	3.21	3.16	3.62	3.27
DIA	0.42	0.48	0.52	0.42	0.48	0.47	0.42	0.43	0.43	0.42

[a] Refer to Table 4.1 for definitions.

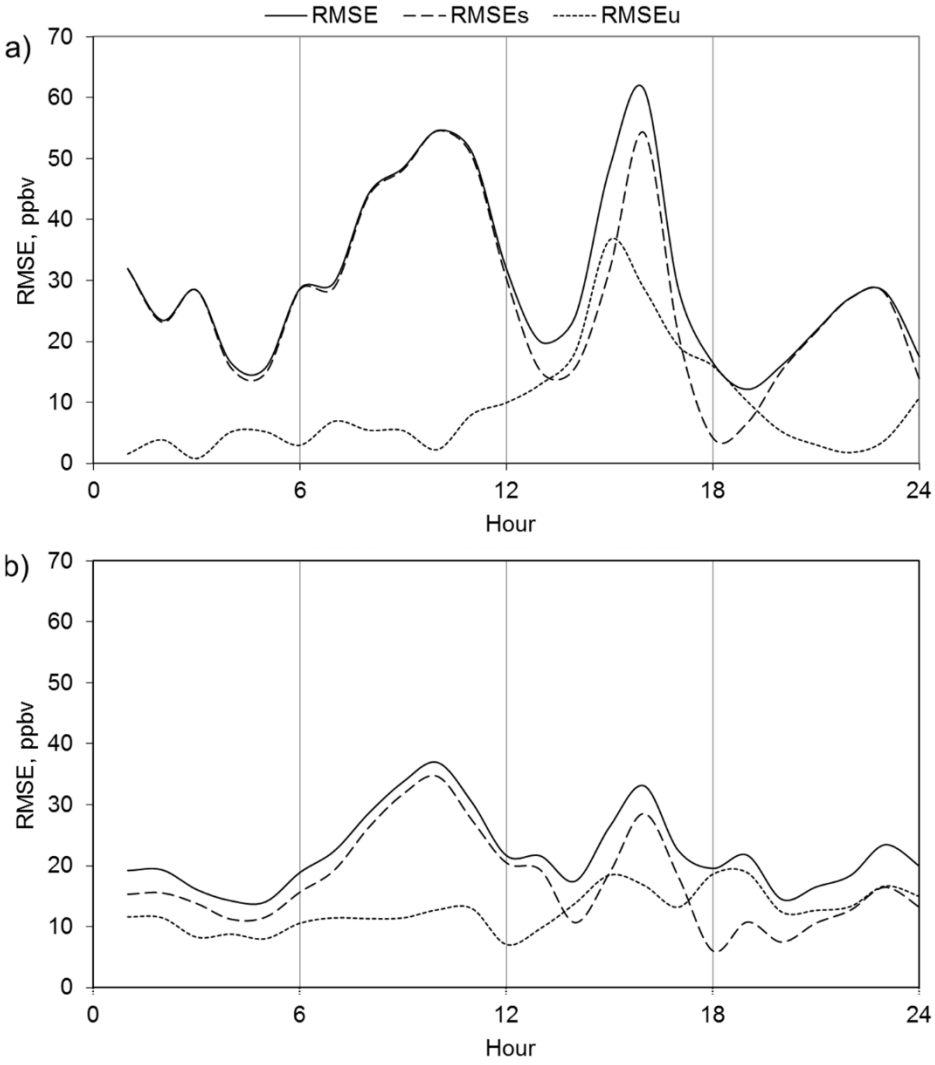

Figure 4.1. Time series for RMSE and its systematic and unsystematic components: a) Base inventory simulation; b) BVLS-corrected inventory simulation.

Because different combinations of regularization techniques and parameter selection methods were assessed, it became evident that it would be useful to evaluate the relative impact of the choice of regularization technique versus the choice of the regularization parameter method. Thus, the results were segregated into clusters of simulations using the same regularization method and the same regularization parameter selection method. Dispersion plots were constructed by pairing observed and simulated species concentrations. Linear regression models were fitted to each of the proposed combinations, which revealed that the use of

the same regularization parameter selection method explained most of the variance shown among the different tested schemes. Figure 4.2 depicts the model response for O_3 using the different regularization parameter selection methods, and Figure 3 presents the model's response for O_3 using different regularization methods. Correlation (R^2) values for regularization schemes using LC were between 0.517 and 0.522, whereas regularization schemes using NCP or GCV for parameter selection ranged between 0.414 and 0.416 and 0.370 and 0.377, respectively. When assembling the results by regularization method (e.g., DSVD or TIKH), DSVD-based methods yielded R^2 values between 0.370 and 0.522, whereas TIKH-based methods yielded values ranging from 0.374 to 0.517. From this analysis, it became clear that R^2 values were more sensitive to changes in the choice of the regularization parameter, regardless of the accompanying regularization method, than to the choice of the regularization method per se.

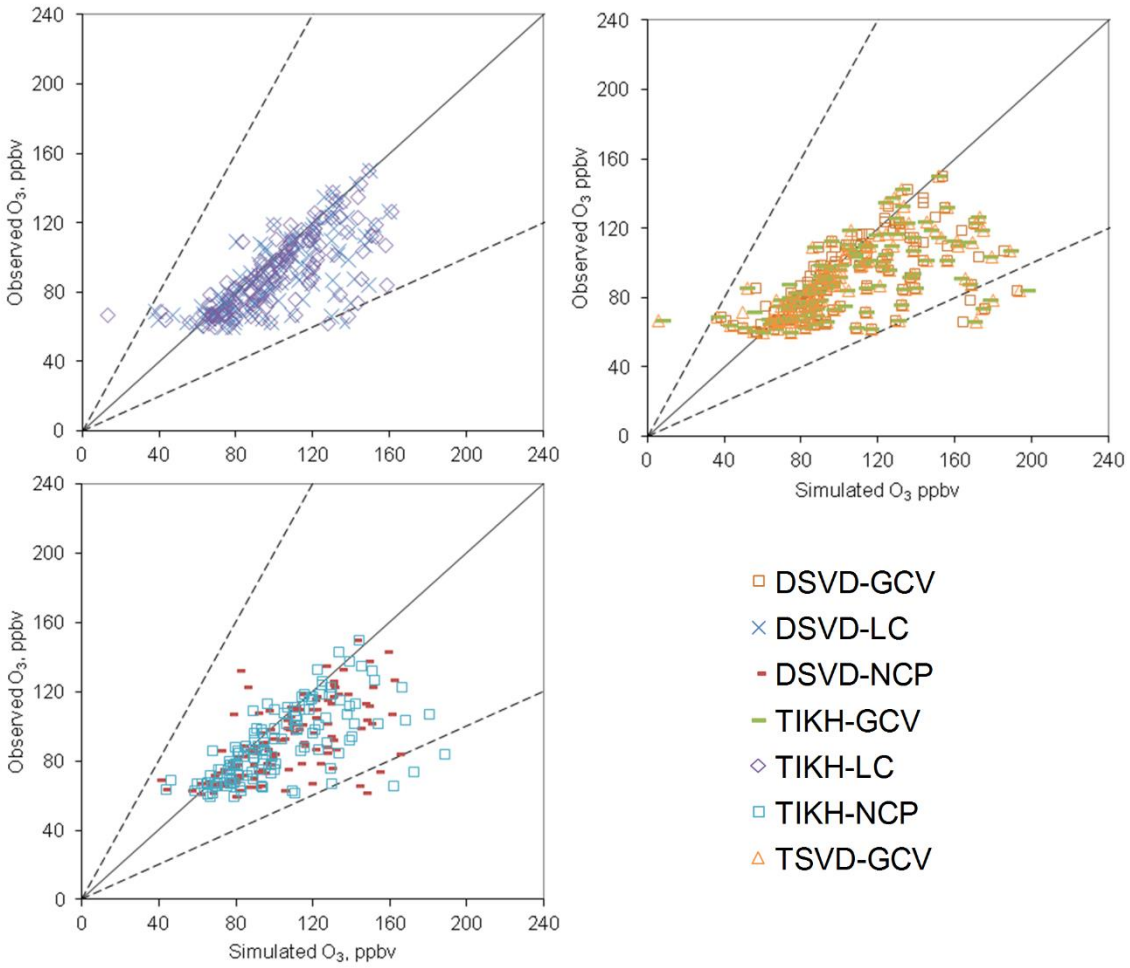

Figure 4.2. Scatter plots for pairs of simulated versus observed O_3 concentrations clustered according to the regularization parameter selection method used: LC (top-right panel), GCV (top-left panel), and NCP (bottom-right panel). Guidelines represent 2:1, 1:1, and 1:2 proportions.

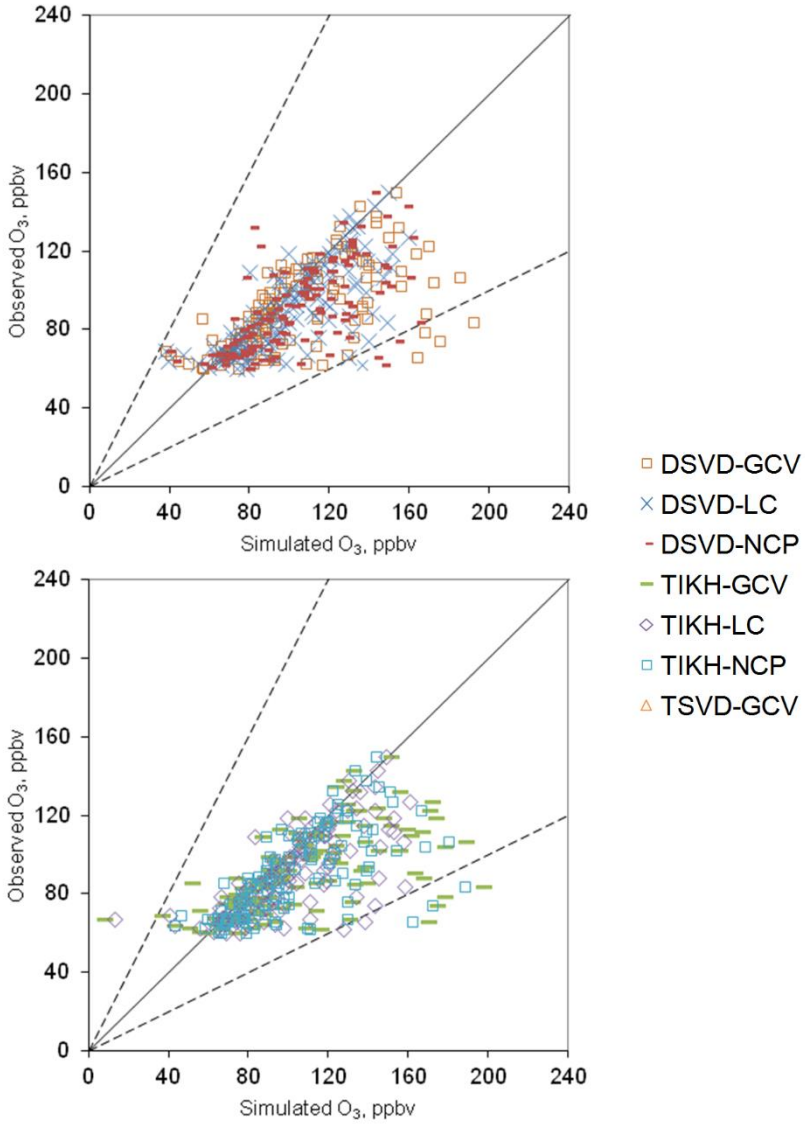

Figure 4.3. Scatter plots for pairs of simulated versus observed O_3 concentrations clustered according to the regularization technique used: DSVD (top panel), TIKH (bottom panel). Guidelines represent 2:1, 1:1, and 1:2 proportions.

Whereas special emphasis has been put on the adequate modeling of ambient O_3 concentrations because of its possible adverse effects on human health and non-linear nature, dispersion plots were also constructed for the species NO_x and CO (Figure 4.4). For these species, as for O_3, most of the regularization methods showed similar behavior among them. This likeness was remarkably evident for CO, as can be seen in Table 4.3. However, as previously shown, differences arose depending on the choice of regularization parameter selection method. For this

particular application, the regularization schemes using the LC method demonstrated better performance than the rest of the schemes.

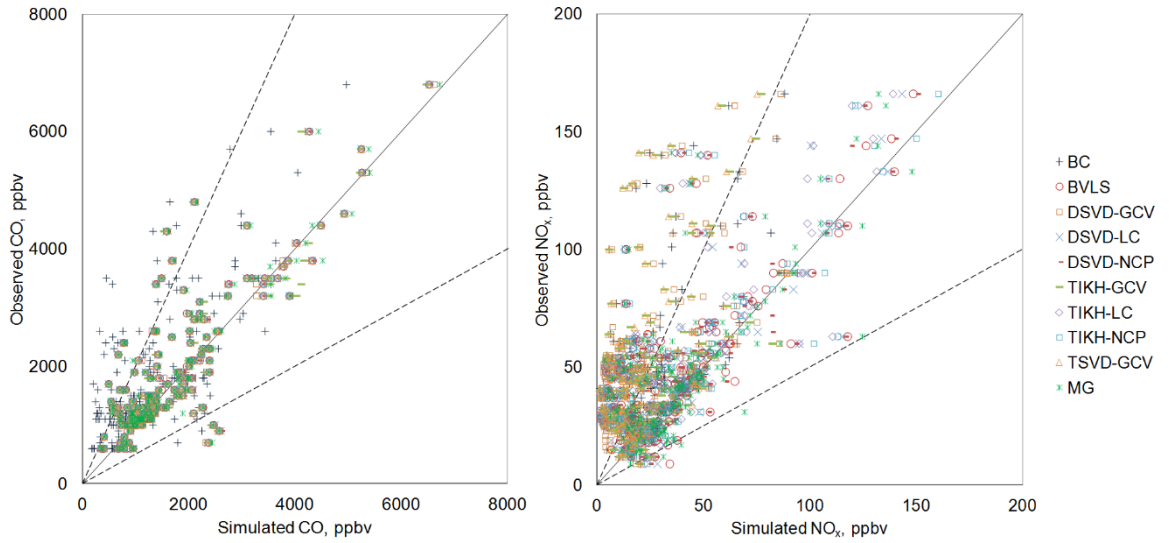

Figure 4.4. Scatter plots for pairs of simulated versus observed concentrations of CO (left panel) and NOx (right panel). Guidelines represent 2:1, 1:1, and 1:2 proportions.

In this study, the correction of NO_x emissions was relevant because within urban environments, O_3 is mostly produced by photochemical reactions between NO_x and volatile organic compounds (VOCs), which explains the high correlation shown by NO_x and O_3 performance metrics for all test runs. That is, correcting the emissions of primary species (NO_x) leads to the automatic improvement in the estimation of secondary species. However, VOC emissions were not corrected because of a lack of proper observations to constrain their inversion. Thus, the remaining differences could be attributed to uncertain VOC emissions.

Finally, after analyzing the behavior of SO_2, it was concluded that this model application, even after undergoing regularization processes, was unable to adequately reproduce the observed SO_2 concentrations within the GMA (Table 4.4). In this regard, it is believed that the level of uncertainty in current SO_2 emission inventories is still large enough as to overcome any effort aimed at correcting these

emissions via mathematical regularization methods. Thus, the spatial distribution of these emissions would need to be revised through a down-top methodology.

In addition to the combinations among regularization methods and regularization parameter-selection methods described above, further combinations are possible. For example, initial exploration of TSVD-LC and TSVD-NCP was conducted. However, they were not further analyzed because similar conclusions on the relevance of the regularization parameter selection method over the regularization technique could be drawn from those initial tests.

4.6 Conclusions

Inverse modeling is being increasingly used as a top-down analysis tool for emissions inventory assessment. Regularization is a mathematical technique that provides numerical stability for this type of ill-conditioned problem and, thus, was explored in this investigation. A common feature to all regularization techniques is that they require the selection of a regularization parameter that seeks to balance the minimization error and the regularization error. In addition, there are restricted least-squares methodologies that solve the inverse problem by restraining the solution to a specific set of boundaries (e.g., BVLS).

In this work, several regularization schemes and one restricted least-squares method were tested and compared using statistical performance criteria with the MG test (2011). This experiment consisted of detecting possible improvements in the emission inventory of ozone precursors at the scale of an urban center (Guadalajara, Mexico, in this case). Overall, the BVLS consistently showed the best agreement among the other mathematical techniques tested. In addition, regularization methods demonstrated almost indistinct behavior patterns among them. Nonetheless, the choice of the regularization parameter was found to explain most of the variance shown among the different tested schemes, with the techniques using the LC method exhibiting better agreement between the observed and simulated values than their NCP and GCV counterparts.

This experiment allowed for the evaluation of the suitability of the use of regularization methods for improving air pollutant emission inventories, and for the direct comparison of a variety of them. The inverse modeling approach was able to significantly reduce However, results also reflected that regularization methods cannot resolve all uncertainties, for example, those related to emission processes of specific pollutants such as SO_2, or the lack of observation data to adequately constrain VOC emissions. Therefore, the need to consider additional approaches and incorporating all available data for a better understanding of pollution sources. In the next Chapter, a supervised machine learning technique -namely, Multivariate Linear Regression (MLR) will be explored as a complementary tool to assess pollutant emission sources.

4.8 References

Allen, D.M., 1974. The relationship between variable selection and prediction. *Technometrics* 16, 125-127.

Aster, R.C., Borchers, B., Thurber, C.H., 2005. *Parameter Estimation and Inverse Problems*, Elsevier Academic Press, Burlington, p. 89-118.

Chai, T., Carmichael, G.R., Tang, Y., Sandu, A., 2009. Regional NOx emission inversion through a four-dimensional variational approach using SCIAMACHY tropospheric NO_2 column observations. *Atmospheric Environment* 43, 5046-5055.

Doll, D.C., Scheffe, R.D., Meyer, E.L., Chu, S.-H., 1991. Guideline for regulatory application of the Urban Airshed Model, EPA-450/4-91-013, Office of Air Quality, Planning and Standards, United States Environmental Protection Agency, Research Triangle Park, 89 pages.

Eckhardt, S., Prata, A.J., Seibert, P., Stebel, K., Stohl, A., 2008. Estimation of the vertical profile of sulfur dioxide injection into the atmosphere by a volcanic eruption using satellite column measurements and inverse transport modeling. *Atmospheric Chemistry and Physics* 8, 3881-3897.

Ekstrom, M.P., Rhodes, R., 1974. On the application of eigenvector expansions to numerical deconvolution. *Journal of Computational Physics* 14, 319-340.

Elbern, H., Strunk, A., Schmidt, H., Talagrand, O., 2007. Emission rate and chemical state estimation by 4-dimensional variational inversion. *Atmospheric Chemistry and Physics* 7, 3749-3769.

Fan, S.-M., Sarmiento, J.L., Gloor, M., Pacala, S.W., 1999. On the use of regularization techniques in the inverse modeling of atmospheric carbon dioxide. *Journal of Geophysical Research* 104(D17), 21503-21512.

Gilliland, A.B., Appel, K.W., Pinder, R.W., Dennis, R.L., 2006. Seasonal NH_3 emissions: Inverse model estimation and evaluation. *Atmospheric Environment* 40, 4986-4998.

Golub, G.H., Heath, M., Wahba, G., 1979. Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics* 21, 215-223.

Haber, E., Oldenburg, D., 1999. A GCV based method for nonlinear ill-posed problems. *Computers and Geosciences* 4, 41-63.

Hakami, A., Henze, D.K., Seinfeld, J.H., Chai, T., Tang, Y., Carmichael, G.R., Sandu, A., 2005. Adjoint inverse modeling of black carbon during the Asian Pacific Regional Aerosol Characterization Experiment. *Journal of Geophysical Research* 110, art. no. D14301.

Hanna, S.R., 1988. Air quality model evaluation and uncertainty. *Journal of Air Pollution Control Association* 38, 460-412.

Hansen, P.C., 1990. Truncated Singular Value Decomposition Solutions to Discrete Ill-Posed Problems with Ill-Determined Numerical Rank. *SIAM Journal on Scientific and Statistical Computing* 11, 503-518.

Hansen, P.C. 1998. Rank-Defficient and Discrete Ill-Posed Problems. Society for Industrial and Applied Mathematics, Philadelphia, pp. 99-131.

Hansen, P.C., O'Leary, D.P., 1993. The use of the L-curve in the regularization of discrete ill-posed problems. *SIAM Journal on Scientific Computing* 14, 1487-1503.

Hansen, P.C., 2008. Regularization Tools. A Matlab Package for Analysis and Solution of Discrete Ill-Posed Problems. Version 4.1 for Matlab 7.4. Technical University of Denmark, pp. 13-42.

Henze, D.K., Seinfeld, J.H., Shindell, D.T., 2009. Inverse modeling and mapping US air quality influences of inorganic $PM_{2.5}$ precursor emissions using the adjoint of GEOS-Chem. *Atmospheric Chemistry and Physics* 9, 5877-5903.

Hoerl, A., Kennard, R.W., 1976. Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics* 12, 55-67.

Krakauer, N.Y., Schneider, T., Randerson, J.T., Olsen, S., 2004. Using generalized cross-validation to select parameters in inversions for regional carbon fluxes. *Geophysical Research Letters* 31, art. no. L19108.

Krawczyk-Stando, D., Rudnicki, M., 2007. Regularization parameter selection in discrete ill-posed problems - The use of the U-Curve. *International Journal of Applied Mathematics and Computer Science* 17, 157-164.

Lawson, C.W., Hanson, R.J., 1974. *Solving Least Squares Problems*. John Wiley and Sons. New York, pp. 350.

Li, M.J., Chen, D.S., Cheng, S.Y., Wang, F., Li, Y., Zhou, Y., Lang, J.L., 2010. Optimizing emission inventory for chemical transport models by using genetic algorithm. *Atmospheric Environment* 44, 3926-3934.

Lin, J., Chen, W., Wang, F., 2011. A new investigation into regularization techniques for the method of fundamental solutions. *Mathematics and Computers in Simulation* 81, 1144-1152.

Meirink, J.F., Bergamaschi, P., Krol, M.C., 2008. Four-dimensional variational data assimilation for inverse modelling of atmospheric methane emissions: method and comparison with synthesis inversion. *Atmospheric Chemistry and Physics* 8, 6341-6353.

Mendoza-Dominguez, A., Russell, A.G., 2001. Emission Strength Validation Using Four-Dimensional Data Assimilation: Application to Primary Aerosol and Precursors to Ozone and Secondary Aerosol. *Journal of the Air & Waste Management Association*, 51, 1538-1550.

Mendoza, A., García, M.R., 2009. Implementation of an air quality model of second generation to the metropolitan area of Guadalajara, Mexico. *Revista Internacional de Contaminación Ambiental* 25, 73-85.

Mendoza, A., García, M.R., 2011. Inverse modeling applied to the analysis of emission inventory of the metropolitan area of Guadalajara, Mexico. *Revista Internacional de Contaminación Ambiental* 27, 199-214.

Miller, C.A., Hidy, G., Hales, J., Kolb, C.E., Werner, A.S., Haneke, B., Parrish, D., Frey, H.C., Rojas-Bracho, L., Deslauriers, M., Pennell, B., Mobley, J.D., 2006. Air emission inventories in North America: a critical assessment. Jo*urnal of the Air & Waste Management Association* 56, 1115-1129.

Napelenok, S. L., Pinder, R.W., Gilliland, A.B., Martin, R.V., 2008. A method for evaluating spatially-resolved NOx emissions using Kalman filter inversion, direct sensitivities, and space-based NO_2 observations. *Atmospheric Chemistry and Physics* 8, 5603-5614.

Neumaier, A., 1998. Solving ill-conditioned and singular linear systems: A tutorial on regularization. *SIAM Review* 40, 636-666.

Pétron, G., Granier, C., Khattatov, B., Lamarque, J.-F., Yudin, V., Muller, J.F., Gille, J., 2002. Inverse modeling of carbon monoxide surface emissions using Climate Monitoring and Diagnostics Laboratory network observations. *Journal of Geophysical Research* 107, ACH 10-1 – ACH 10-23.

Quélo, D., Mallet, V., Sportisse, B., 2005. Inverse modeling of NOx emissions at regional scale over Northern France. Preliminary Investigation of the Second Order Sensitivity. *Journal of Geophysical Research* 110, art. no. D24310.

Russell, A., Dennis, R., 2000. NARSTO critical review of photochemical models and modeling. *Atmospheric Environment* 34, 2283-2324.

Russell, A.G., 2008. EPA Supersites program-related emissions based particulate matter modeling: initial applications and advances. *Journal of the Air & Waste Management Association* 58, 289-302.

Rust, B.W., 2000. Parameter selection for constrained solutions to ill-posed problems. *Computing Science and Statistics* 32, 333-347.

Rust, B.W., O'Leary, D.P., 2008. Residual Periodograms for Choosing Regularization Parameters for Ill-Posed Problems. *Inverse Problems* 24, 1-30.

Saide, P., Osses, A., Gallardo, L., Osses, M., 2009. Adjoint inverse modeling of a CO emission inventory at the city scale: Santiago de Chile's case. *Atmospheric Chemistry and Physics* 9, 6325-6361.

Stark, P.B., Parker, R.L., 1995. Bounded-Variable Least-Squares: an Algorithm and Applications. *Computational Statistics* 10, 129-141.

Stohl, A., Seibert, P., Arduini, J., Eckhardt, S., Fraser, P., Greally, B. R., Lunder, C., Maione, M., Mühle, J., O'Doherty, S., Prinn, R. G., Reimann, S., Saito, T., Schmidbauer, N., Simmonds, P. G., Vollmer, M. K., Weiss, R. F., Yokouchi, Y., 2009. An analytical inversion method for determining regional and global emissions of greenhouse gases: Sensitivity studies and application to halocarbons. *Atmospheric Chemistry and Physics* 9, 1597-1620.

Tian, D., Cohan, D.S., Napelenok, S., Bergin, M., Hu, Y., Chang, M., Russell, A.G., 2010. Uncertainty analysis of ozone-formation and response to emission controls using higher-order sensitivities. *Journal of the Air & Waste Management Association* 60, 797-804.

Winiarek, V., Bocquet, M., Saunier, O., Mathieu, A., 2012. Estimation of errors in the inverse modeling of accidental release of atmospheric pollutant: Application to the reconstruction of the cesium-137 and iodine-131 source terms from the Fukushima Daiichi power plant. *Journal of Geophysical Research* 117, art. no. D05122.

Xiao, X., Prinn, R.G., Fraser, P.J., Weiss, R.F., Simmonds, P.G., O'Doherty, S., Miller, B.R., Salameh, P. K., Harth, C.M., Krummel, P.B., Golombek, A., Porter, L.W., Butler, J.H., Elkins, J.W., Dutton, G.S., Hall, B.D., Steele, L.P., Wang, R.H.J., Cunnold, D.M., 2010. Atmospheric three-dimensional inverse modeling of regional industrial emissions and global oceanic uptake of carbon tetrachloride. *Atmospheric Chemistry and Physics* 10, 10421-10434.

Yang, Y.-J., Wilkinson, J.G., Russell, A.G., 1997. Fast, direct sensitivity analysis of multidimensional photochemical models. *Environmental Science and Technology* 31, 2859-2868.

Zuk, M., Rojas Bracho, L., Tzintzun Cervantes, M.G., 2007. *Tercer almanaque de datos y tendencias de la calidad del aire en nueve ciudades mexicanas*. Instituto Nacional de Ecología, Mexico, pp. 116.

This page intentionally left blank

Chapter 5

Impacts of economic and sociodemographic variables on air quality in Mexican Metropolitan Areas

This chapter was coauthored with Sergio Santiago Cárdenas-Pérez, Diego Guillermo González-Almendárez, Isabel Cristina Xicoténcatl-Guzmán, and Alberto Mendoza-Domínguez.

As explained in the preceding chapters, while deterministic models are very powerful to resolve complex atmospheric emission-concentration processes, one of their shortcomings is the need for detailed emission inventories, that even nowadays carry large uncertainties which the use of mathematical tools cannot fully overcome. In this chapter, supervised machine learning methods -specifically, multiple linear regression models- are explored as complementary tools for a better understanding of pollution drivers in Mexican cities.

5.1 Introduction

Ambient air pollution is considered the greatest environmental threat to human health. Almost the entire global population (99%) breathes unhealthy levels of fine particulate matter and nitrogen dioxide. People in low and middle-income countries are most exposed to outdoor air pollution at levels exceeding World Health Organization (WHO) air quality limits. Outdoor and indoor air pollution were responsible for approximately 7 million deaths globally in 2016. The highest attributable mortality rates were concentrated in lower-middle-income countries that accounted for almost nine of 10 (88%) deaths (WHO, 2022a; WHO, 2022b).

Metropolitan areas are highly complex, diversified spaces, in which different urban functions -administrative-political, productive, commercial, housing, cultural, recreational, and touristic- are carried out simultaneously (Ramírez Sáiz and Safa Barraza, 2011). Mexican metropolitan areas are defined as groups of municipalities

that interact around a major city with more than 50,000 inhabitants (Ortega-Montoya et al. 2021). Their rapid growth has spawned several social and environmental issues, including urban air pollution. The three major Mexican metropolitan areas: Mexico City Metropolitan Area (MCMA), Monterrey Metropolitan Area (MMA) and Guadalajara Metropolitan Area (GMA) are not exempted (INECC, 2021) of this problem.

In this context, air quality models (AQM) are mathematical modeling tools that allow studying the dynamics of atmospheric pollutants and can be useful to study emitter-receptor relationships, forecast poor air quality events, and eventually establish public policy instruments for their abatement. Deterministic AQM solve the primitive conservation equations (matter, energy, momentum and chemical species), combined with parameterizations of complex atmospheric phenomena (e.g., turbulent convection, soil-atmosphere interaction, cloud microphysics, etc.) to describe states of the atmosphere and the evolution of pollutants) (Kadaverugu et al., 2019). However, the application of such models requires considerable efforts to generate the required information and to computationally solve the modeled scenarios. (See Chapters 1-4 of this dissertation).

With the increasing availability of large amounts of historical data, deterministic models are being complemented with empirical models to answer questions related to air pollution. For example, Carmona et al. (2020, 2021) has combined satellite data with Artificial Neural Networks (ANN) to study the levels of fine suspended particles in Northeast Mexico. Iglesias-Gonzalez et al. (2020) forecasted systems based on time series models. Within machine learning techniques, in the family of statistical models, multiple linear regression models (MLRM) are receiving increased attention for their application to environmental problems. MLRM can be used to explore and quantify the interactions and contributions of socioeconomic systems to observed air pollution levels using historical sectoral indicators, and thus help governments to diagnose, monitor, and forecast macroeconomic outcomes to reduce or maintain allowable emissions and

ensure effective public and environmental policies (Shpak et al. 2022). In the literature, multiple applications of MLR models to study air quality have been documented, e.g. Rosenlund et al. (2008) Ganesh et al. (2017), Bai et al. (2018), Ganesh et al. (2019), Abdullah et al. (2020), Shams et al. (2021), He et al. (2022), the majority of them considering meteorological and historical pollutant data as predictor variables.

Econometric science has been long employed to study the impact of population and macroeconomic indicators on the levels of atmospheric carbon dio CO_2 emissions in the United States and the Asia-Pacific region. Cramer (1998) produced one of the first works to examine the impact of population growth on air pollution in California and conclude that population is closely associated with some sources of emissions but not with others. Later, by considering sulfur dioxide (SO_2), Cole and Neumeyer (2004) presented the first work that explicitly examined the impact of demographic factors on a pollutant other than carbon dioxide at the cross-national level, and took into account the urbanization rate and the average household size neglected by many prior cross-national econometric studies. For carbon dioxide emissions, Cole and Neumeyer (2004) found evidence that population increases are matched by proportional increases in emissions while a higher urbanization rate and lower average household size increase emissions. For sulfur dioxide emissions, they found a U-shaped relationship, with the population-emissions elasticity rising at higher population levels. Urbanization and average household size are not found to be significant determinants of sulfur dioxide emissions. For both pollutants, our results suggest that an increasing share of global emissions will be accounted for by developing countries (Cole and Neumeyer, 2004).

More recently, the method of a correlation-regression analysis with the subsequent construction of econometric levels was utilized by Shpak et al. (2022) confirming the dependence of CO_2 emissions on GDP, exports and imports, the rate of inflation and unemployment in the United States. The influence of socioeconomic

variables has also been studied earlier in China. Yang et al. (2022) developed a geographically weighted regression (GWR) model and concluded that air quality was more sensitive to variations in socioeconomic metrics in less developed and medium-sized cities. Zhou, Li, and Zhang (2022) studied the effect of meteorological conditions, economic features, and population density due to variations in marine and terrestrial geographical environments on air quality through the application of a multiscale GWR model and wavelet analysis. They showed that the relationship between population density and urbanization rate with ozone concentration was greater in coastal cities, and that daily maximum temperature was the most important factor influencing O_3 concentration.

In Latin America, socioeconomic information, such as income, multidimensional, and energy poverty levels, has been used along ordinary least squares regression models to determine the relationship between pollution and socioeconomic information in Chilean cities. The obtained models showed positive and significant correlations between income, multidimensional, and energy poverty and the different pollutants, but mixed results in the case of unemployment (Herrera, Rojo, and Scapini, 2022)..

Although MCMA, MMA and GMA are distinctly characterized by different geographic and meteorological conditions (Stoltz et al. 2020), there has been some research aiming to compare their air quality characteristics. Hernández-Paniagua et al. (2017) performed an assessment of long-term trends in O_3 and odd oxygen ($O_3 + NO_2$) at MMA and compared it to MCMA and GMA. Ramírez Sáiz and Safa Barraza (2011) provided an overview of the urban conditions in the three metropolitan areas in different time periods but did not address explicitly the problem of bad air quality. Ortega-Montoya et al. (2021) analyzed the spatial configuration of accidental chemical risk scenarios in MCMA, GMA and MMA, and performed complementary geostatistical correlation analyses using population data, marginalization indexes, and industrial clustering sectors to identify trends that would lead to environmental justice approaches. Stoltz et al. (2020) analyzed

the spatial representativeness of air quality stations through the use of clustering techniques, showing that GMA and MMA have a well distributed air quality network with the fewest number of similar stations, while in contrast, in the MCMA, a cluster of possible redundant stations is found. Understanding the differences and similitudes in these three metropolitan areas can allow for a better design and implementation of pollution control measures and public policies aimed at reducing air pollution and improve people's welfare.

This research investigates the regional air quality characteristics and its drivers in MCMA, MMA and GMA. The specific objectives are to identify long-term, economic and sociodemographic indicators, routinely reported by Mexican government agencies, that might have an influence in air pollution in the selected metropolitan areas, and to build and assess MLRM based on economic and sociodemographic variables to predict air quality in Mexican metropolitan that allow for the quantification of urban air pollution causes in these regions. To the authors' knowledge, it is the first study to explore and compare for MCMA, MMA and GMA the relationship between economic and sociodemographic variables, only, and air pollution, by using monthly, long-term, publicly available government data.

5.2 Methods

5.2.1 Area of study

5.2.1.1 Mexico City Metropolitan Area

The MCMA comprises Mexico City and 60 adjoining municipalities, with an area of 5954 km^2. The MCMA is located in an endorheic lake basin with an average altitude of 2250 m a.s.l., which is surrounded at the north by Sierra de Guadalupe, at the southeast by Santa Catarina Mountain, at the south by Sierra de Chihinautzin, and at the east and west by Sierra Nevada and Sierra de las Cruces, respectively (INECC, 2007). It is the most populated area in Mexico, as well as the

seventh largest megacity in the world (UN, 2016) with 22 million inhabitants (INEGI, 2020), equivalent to 17% of the country's total population.

The main economic activities in MCMA are commerce, financial and insurance services, transport, and tourism. The MCMA creates around 23% of the national Gross Domestic Product (GDP), making it an extremely important part of Mexico's economy. The primary land use is for urban settlements, but there is also a large area of vegetation in the outskirts of this region. In 2018, there were 6 million vehicles in the MCMA, of which 89.4% where private-owned, 5.1% freight, and 5.5% public transportation. Due its large size, air pollution remains an important issue in MCMA. For example, in 2020, the number of days exceeding the maximum permissible level for at least one pollutant was 262 (72%). The most frequently exceeded pollutant was O_3 (64% of the days), followed by PM_{10} (21%), SO_2 (12%) and $PM_{2.5}$ (4%) (INECC, 2021).

5.2.1.2 Monterrey Metropolitan Area

The MMA in northeast Mexico, is formed by Monterrey City and 17 municipalities: Abasolo, Apodaca, Cadereyta Jiménez, El Carmen, Ciénega de Flores, García, San Pedro Garza García, General Escobedo, General Zuazua, Guadalupe, Juárez, Monterrey, Pesquería, Salinas Victoria, San Nicolás de los Garza, Hidalgo, Santa Catarina and Santiago. According to INEGI (2020), MMA is the second-most populated area in Mexico with a population of 5,341,177 inhabitants within a 7657 km^2 area. The MMA lies an open plain with an average altitude of 530 m a.s.l., crossed by the Pesqueria river at the north and Santa Catarina River at the south. The MMA is surrounded by Sierra Madre Oriental at the south, which has an altitude up to 2400 m a.s.l. and Sierra de la Silla with elevations between 1200 and 1800 m a.s.l.

The main economic activities in the MMA are services and manufacturing. The city's urban mobility statistics have shown an increasing trend in the number of vehicles, while the use of public transportation has been decreasing. According to

Instituto de Control Vehicular del Estado de Nuevo León, in 2021 there were 2,587,209 registered vehicles in the State of Nuevo León, an increase of 44% with respect to the number of vehicles registered in 2011 (1,792,905 vehicles). The state of Nuevo Leon has a GDP per capita of 18,912, 88% larger than the national average (Datos Nuevo Leon, s.f.). According to SEMARNAT (2016), the state uses 2,967.42 kWh of electricity per home, 0.30 toe of fuel per home, and 0.55 toe of total energy per home.

In 2020, the number of days exceeding the maximum permissible level for at least one pollutant within the MMA was 207, which represented an increase from 2019 (188 days), but still lower than the previous 20 years (which ranged from 223 days in 2018 to 320 days in 2011). The most frequently exceeded pollutant is PM_{10} (45% of the days), followed by O_3 (22%) and $PM_{2.5}$ (8%) (INECC, 2021).

5.2.1.3 Guadalajara Metropolitan Area

The GMA is in central Mexico and comprises Guadalajara City and 10 municipalities with 5.2 million inhabitants in an area of 2735 km^2. The GMA is located in the Valley of Atemajac, approximately 500 km northwest of the MCMA, and an altitude of 1600 m a.s.l. In contrast to MCMA and MMA, GMA is not surrounded by mountains, except at the northeast where Barrancas de Oblatos canyon is located, which a depth of 600 m and a maximum altitude of 1520 m a.s.l.

In the GMA, the main economic activities are industry and services (Ortega-Montoya et al. 2021). The state of Jalisco is the fourth largest contributor to the national GDP, with 6.9% of the total. The GMA concentrates 64.3% of the total vehicle fleet in the state of Jalisco, with 2,514,679 vehicles in circulation, and consumes 60% of the state's energy consumption. Jalisco is also the fourth largest consumer of energy at the national level. In GMA, in 2020, the number of days exceeding the maximum permissible level for at least one pollutant was 159, a decrease from the 267 exceedance-days reported in 2019. The most-frequent exceeded pollutant is PM_{10}, followed by O_3. Although average PM concentration

has shown a decreasing trend during the period 2016-2020, still, in 2020, 25% of the days exceeded the daily average standard for PM_{10}. Other pollutants (e.g. $PM_{2.5}$ and O_3) have remained relatively stable (INECC, 2021).

Figure 5.1 shows the location of MCMA, MMA and GMA, including land use, territorial division, and location of air quality network (Benítez-García et al. 2014).

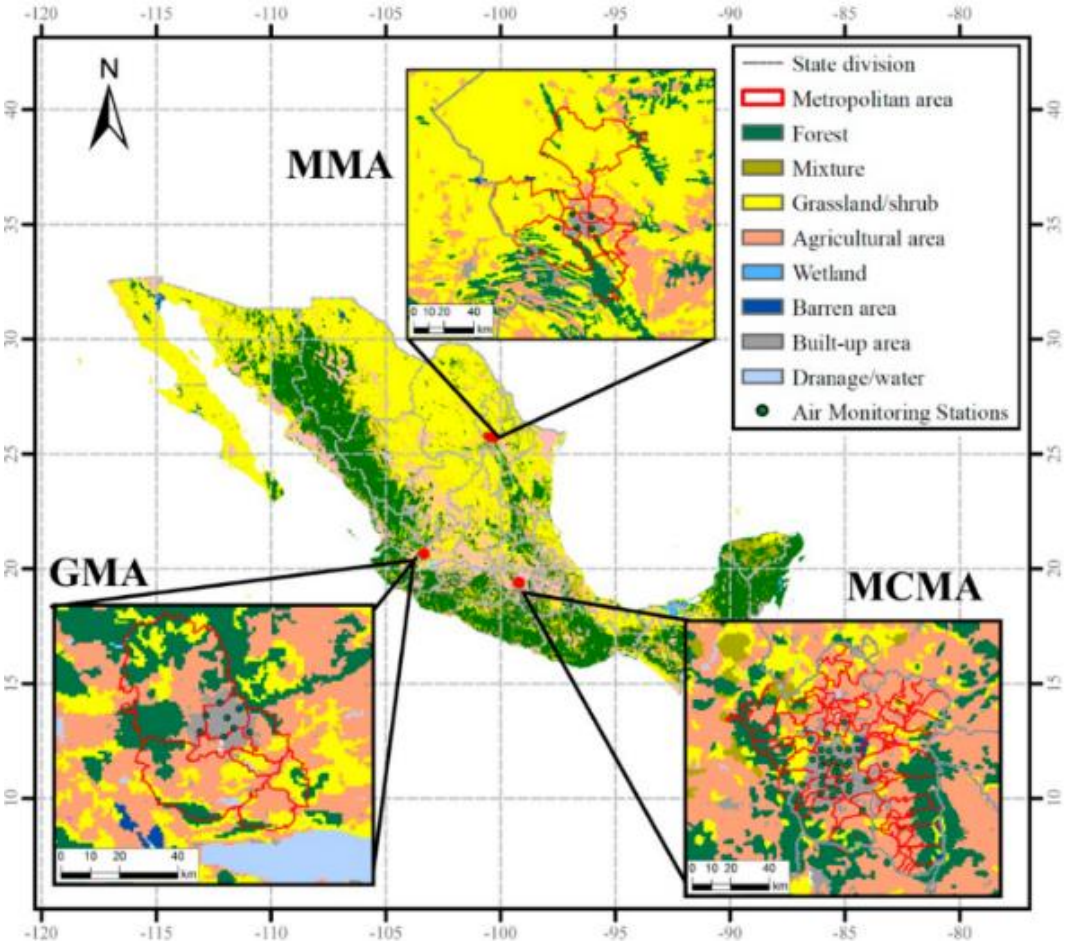

Figure 5.1 Locations of the three Mexican metropolitan areas: MCMA, MMA and GMA, their land use, territorial division, and locations of the air-monitoring stations. (Taken from: Benitez-García et al. 2014)

5.2.2 Model philosophy

It is known that pollutant emissions are a direct consequence of human activities, for example, commercial activities typically involve transportation, which inevitably produce pollutant emissions (e.g. CO_2, CO, NO_x, PM_{10}, and $PM_{2.5}$) as a result of the incomplete combustion of fuels. Under this context, it can be expected that having data on the use of fuels can provide an overview of pollution dynamics in a certain geographic region. The MLR model proposed here relies on this premise by correlating long-term economic and energy indicators, routinely reported on a monthly basis by Mexican government agencies (e.g. INEGI, Secretaría de Energía), with monthly-averaged air pollution data for the MMA, GMA and MCMA. Since air quality is the product of complex interactions between meteorological variables, geographical factors, atmospheric processes and emissions sources, socioeconomic variables might not explain all variance in the observed pollutant concentrations, but can allow the identification and analysis of activities with impact in air quality. Figure 5.2 depicts the methodology framework followed in this study.

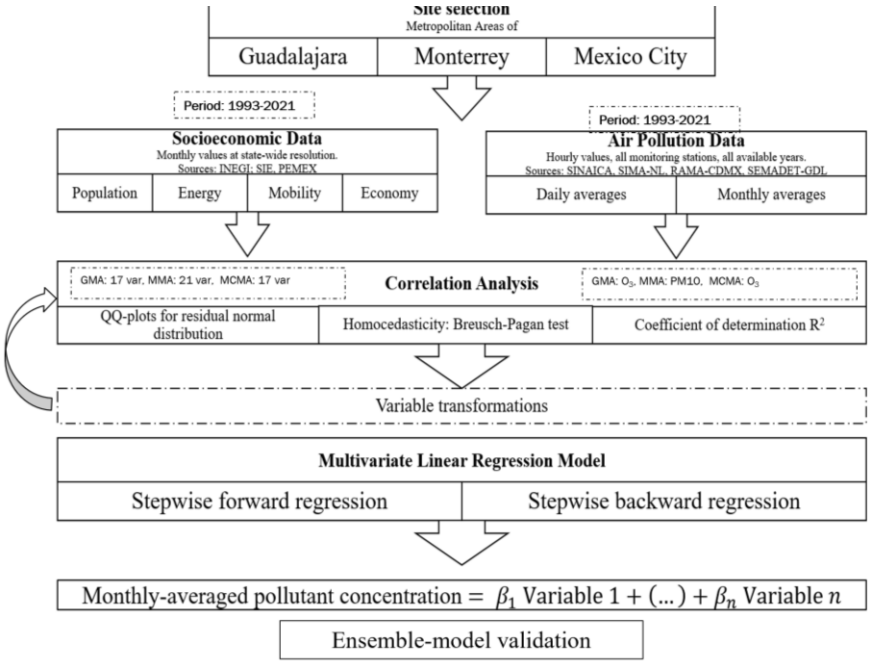

Figure 5.2 Methodology framework.

5.2.3. Databases

5.2.3.1 Air quality data and pollutant selection

Long-term hourly measurements of criteria air pollutants were extracted from records of the MCMA, MMA, and GMA official air quality networks, namely Red Automática de Monitoreo Atmosférico (RAMA), Sistema Integral de Monitoreo Ambiental (SIMA) and Sistema de Monitoreo Atmosférico de Jalisco, respectively. Data quality in monitoring sites is assured by the measurement methods and calibration procedures defined by the Mexican Secretariat for the Environment and Natural Resources (SEMARNAT, 2011). Mexican standard NOM-156-SEMARNAT-2012 describes maintenance procedures, quality assurance, and quality control processes that must be followed to secure the traceability and continuous quality of air monitoring data.

To narrow the scope of this research, only one pollutant was selected as proxy for pollution levels in each metropolitan area. This pollutant was chosen based on data quality (i.e., no evidence of drifting or shifting in time series records) and data availability (≥75% per year in agreement with Mexican standards). For example, in the case of MCMA RAMA sites, if a monitoring site displayed a three-year gap between data measurements, was removed prior year 2018, or had an operation span of less than 3 years, the site will not be considered in this research. Also, because monitoring data is reported in hourly basis, daily and monthly averages were computed. As a result, the selected pollutants were PM_{10} for MMA, and O_3 for MCMA and GMA. Of note, PM_{10} and O_3 also typically exceed the air quality standards in said Mexican metropolitan areas (Soltz et al. 2020, INECC, 2021).

For each pollutant, all distinct monitoring sites follow similar temporal trends, from which one can hypothesize that the average of all monitoring sites within a domain can adequately represent the pollution levels in the domain. In the case of MMA, previous works have demonstrated that, for $PM_{2.5}$, SW, CE, and NE stations are spatially similar to one another (Mancilla et al. 2019). Moreover, a recent

assessment of the spatial representativeness of air quality monitoring networks in the MMA, MCMA and GMA concluded that GMA has a well distributed air quality network with the fewest number of similar stations, as well as the MMA, which presents the same stations clusters for PM_{10} and O_3. In contrast, in the MCMA, a cluster of possible redundant stations is found (Stolz et al. 2020). The suitability of computing a "representative domain-average" is further confirmed by analyzing monthly-averaged concentration boxplots per site.

5.2.3.2 Economic and socio-demographic data

A total of 107 variables with potential relation to air quality were identified from the public, official sites of Instituto Nacional de Estadística y Geografía (INEGI) and Sistema de Información Energética (SIE) databases. The variables could be classified in energy, population, economy and mobility categories. To further correlate them with air pollution data, only those variables with monthly-updated, long-term records, and statewide or metropolitan geographical resolution where opted for, and others discarded.

In the case of MCMA, 83 variables from several categories -ranging from building construction indices, personnel working hours, worker wages, to fuel production and number of registered vehicles- met the selection criteria to be further correlated to monthly-averaged O_3 concentrations. In the cases of MMA and GMA, 55 and 33 variables, were also selected as potential predictor variables to be correlated through univariate linear regression models to monthly-averaged PM_{10} and O_3, respectively.

5.2.4 Statistical analysis

Descriptive statistics (maximum, minimum, mean, median, standard deviation) were computed for all variables. Linear regressions for the selected variables were evaluated in terms of their coefficient of determination (R^2), homoscedasticity, and significance (p-value). R^2 represents the proportion of the variation in the dependent variable that is predictable from the independent variables. According to

literature, R^2 >0.08 can be considered significant (Zhao, 2022). The existence of heteroscedasticity is a major concern in regression analysis and the analysis of variance, as it invalidates statistical tests of significance that assume that the modelling errors all have the same variance. Here, the Breusch-Pagan test with a level of significance of 0.05 was used to validate the homoscedasticity of linear regressions. Due to the different nature of the available variables and with the aim to increase the number of suitable variables for multiple regression analysis purposes, all variables were transformed and plotted against a transformation (e.g. logarithm, square root, inverse, squared inverse, squared, cubed form) of the corresponding monthly-averaged pollutant concentration.) to increase the regression coefficient or assure homoscedasticity.

5.2.4.1 Multiple linear regression model

Multiple-linear regression (MLR) models relate a set of predictor variables via some linear function to a predictand variable (Maraun and Widmann, 2018), and are given as:

$$\boldsymbol{y = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_i x_i + e_i} \quad \textbf{(5.1)}$$

where $\boldsymbol{y}$ is the dependent variable, $\boldsymbol{x}$ and $\boldsymbol{x_i}$ are the independent variables, $\boldsymbol{b}$ and $\boldsymbol{b_i}$are the regression coefficients, and $\boldsymbol{e}$ is the error. Stepwise regression is the step-by-step iterative construction of a regression model that involves the selection of independent variables to be used in a final model. It involves iteratively adding or removing potential explanatory variables in succession and testing for statistical significance. The forward selection approach starts with nothing and adds each new variable incrementally, testing for statistical significance. The backward elimination method begins with a full model loaded with several variables and then removes one variable to test its importance relative to overall results.

After narrowing down the set of predictor variables, stepwise (backward and forward) MLR models were computed using the 'olsrr', 'gvlma', 'car', 'lmtest'

packages in R Studio v4.2.1. Here, it was noted that if the dependent variable was reexpressed, such reexpression had had to be applied to all variables, so models for linear (no transformation), squared, and inverse distributions were created. The resulting MLR models performance was evaluated in terms of significance (p-value), homoscedasticity (Breusch-Pagan test), and the F-statistic, which describes the ratio of two variances, i.e., two mean squares. Mean squares are simply variances that account for the degrees of freedom (DF) used to estimate the variance. Additionally, residuals autocorrelation was tested through the Durbin Watson (DW) statistic. DW will always have a value ranging between 0 and 4, in which a value of 2.0 indicates there is no autocorrelation detected in the sample. Values from 0 to less than 2 point to positive autocorrelation and values from 2 to 4 means negative autocorrelation.

Cook´s Distance and QQ (Quantile-Quantile) plots were also analyzed. Cook's distance plot finds influential outliers in a set of predictor variables, and can help identify points that negatively affect the regression model. The measurement is a combination of each observation's leverage and residual values, the higher the leverage and residuals, the higher the Cook's distance. QQ plots, on the other hand, are probability plots that graphically compare two probability distributions by plotting their quantiles against each other. A point on the plot corresponds to one of the quantiles of the second distribution plotted against the same quantile of the first distribution.

5.3 Results and discussion

5.3.1 MLR model for O_3 in the MCMA

Table 5.2 presents the list of the evaluated INEGI-SIE variables included in the correlation analysis per metropolitan area. In the case of the MCMA, the 83 identified variables were narrowed to 17 variables. All 17 variables showed R^2 >0.05. These variables include MCMA data, but also include information from the the Miguel Hidalgo Refinery, located approximately 87 km northwest of Mexico City

in the state of Hidalgo, whose operations have been documented to have an effect in MCMA (García-Escalante et al., 2014; Sosa et al., 2013). Table 5.1 presents descriptive statistics and general information for energy-related (refinery) variables considered for the MCMA, while Table 5.2 displays the rest of the selected socioeconomic variables for the MCMA. Tables 5.3 and 5.4 present the univariate model performance statistics, indicating if any re-expression was performed to increase the regression coefficient or assuring homocedasticity.

Stepwise forward and backward regressions were computed, initially considering all 17 variables, but also subsets were analyzed to further investigate in the model´s statistical performance and significance of variables. Regarding stepwise regression, it is recommended that the variance inflation factor (VIF) to be below 3 or 4, and the tolerances to be above 0.25 for a robust model. Although the R^2 is an important metric, VIF and tolerances must also be carefully considered. It was noticed that while most stepwise models yielded similar results, the best performance was yielded by a forward-stepwise model showing better increased tolerances and low VIF. Effort was also placed in obtaining the most simple model with the best performance statistics.

Table 5.1 Refinery-related selected socioeconomic variables for the MCMA.

Socioeconomic variables					Descriptive Statistics			
Predictor Variable ID	**Variable description**	**Category**	**Data source**	**Data availability (spatial, temporal)**	**Min**	**Mean**	**Max**	**Standard Deviation**
E.1	Crude oil and liquids process in refinery	Refinery	SIE	Monthly data from 1993 to 2021 Sectioned by Refinery	521.7939	176.7229	643.746	108.4712
E.1.1	Crude oil and liquids process in refinery (Heavy)	Refinery	SIE	Monthly data from 1993 to 2021 Sectioned by Refinery	136.9867	40.4938	223.602	43.8033
E.2	Production of petroleum products by refinery	Refinery	SIE	Monthly data from 1993 to 2021 Sectioned by Refinery	532.3886	202.3486	722.914	111.3417
E.2.2	Production of petroleum products by refinery (Liquid gas)	Refinery	SIE	Monthly data from 1993 to 2021 Sectioned by Refinery	19.5791	4.9083	36.752	6.2269
E.2.5.1	Production of petroleum products by refinery (Pemex Diesel)	Refinery	SIE	Monthly data from 1994 to 2021 Sectioned by Refinery	93.856	14.6224	162.53	34.2124
E.2.6	Production of petroleum products by refinery (Fuel Oil)	Refinery	SIE	Monthly data from 1993 to 2021 Sectioned by Refinery	163.875	63.6595	228.488	33.887

Table 5.2 Non-refinery related selected socioeconomic variables for the MCMA.

Socioeconomic variables					Descriptive Statistics			
Predictor Variable ID	**Variable description**	**Category**	**Data source**	**Data availability (spatial, temporal)**	**Min**	**Mean**	**Max**	**Standard Deviation**
E.3	Internal demand for Gas	Gas and Gasoline	SIE	Monthly data from 1995 to 2019 Sectioned by State	80.2531	65.42092	99.8388	7.7024
E.4.1	Volume of internal sales of Gasoline by federal entity	Gas and Gasoline	SIE	Monthly data from 1993 to 2020 Sectioned by State	143.0519	116.4964	165.6005	13.3536
E.4.1.2	Volume of internal sales of Gasoline by federal entity (Pemex Magna)	Gas and Gasoline	SIE	Monthly data from 1993 to 2020 Sectioned by State	580.0607	232.0056	754.5108	109.0959
E.4.3	Volume of internal sales of Diesel by federal entity	Gas and Gasoline	SIE	Monthly data from 1993 to 2020 Sectioned by State	38.5175	22.31879	46.8837	5.3752
E.5	Internal Demand for Natural Gas by State	Gas and Gasoline	SIE	Monthly data from 1993 to 2019 Sectioned by State	353.9687	139.4123	584.3652	91.7863
E.6	Internal Demand for Natural Gas by State, Public Sector	Gas and Gasoline	SIE	Monthly data from 1993 to 2019 Sectioned by State	178.7571	77.89788	328.5038	41.0776
E.8	Internal Demand for Natural Gas by State, Industrial Sector	Gas and Gasoline	SIE	Monthly data from 1993 to 2019 Sectioned by State	22.3154	2.75462	44.6849	9.743
E.16	Registered motor vehicles in circulation	Mobility	INEGI	Monthly data from 1993 to 2021 Sectioned by State	5.5001	2.901888	8.9704	1.9184
E.16.1	Registered motor vehicles in circulation (Cars)	Mobility	INEGI	Monthly data from 1993 to 2021 Sectioned by State	4.5999	2.33168	7.7995	1.7065
E.16.2	Registered motor vehicles in circulation (Buses)	Mobility	INEGI	Monthly data from 1993 to 2021 Sectioned by State	34.3767	13.456	44.582	10.4892
E.16.3	Registered motor vehicles in circulation (Loading Vehicles)	Mobility	INEGI	Monthly data from 1993 to 2021 Sectioned by State	865.8339	512.366	1138.113	213.3087

Table 5.3 Univariate model performance statistics for MCMA selected predictor variables (refinery-related).

Univariate regression model						
Predictor Variable ID	**R^2**	**P value**	**Homocedasticity test (Brreusch Pragan test)**	**Independent variable scaling**	**Independent variable re-expression (function)**	**Dependent variable re- expression (function)**
E.1	0.0652	3.2553e-06	BP = 3.841761 df = 1 p-value = 0.049991	None	x^2	x^2
E.1.1	0.094	1.7494e-08	BP = 4.228745 df = 1 p-value = 0.039745	None	$1/x^2$	x^2
E.2	0.0733	7.4771e-07	BP = 4.497231 df = 1 p-value = 0.03395	None	x^2	x^2
E.2.2	0.0801	2.2157e-07	BP = 9.11003 df = 1 p-value = 0.002542	None	None	x^2
E.2.5.1	0.081	2.1352e-07	BP = 8.676032 df = 1 p-value = 0.003224	None	$1/ x^2$	x^2
E.2.6	0.1461	1.0388e-12	BP = 6.467843 df = 1 p-value = 0.010984	None	None	X^3

Table 5.4 Univariate model performance statistics for MCMA selected predictor variables (non-refinery-related).

Univariate regression model						
Predictor Variable ID	**R^2**	**P value**	**Homocedasticity test (Brreusch Pragan test)**	**Independent variable scaling**	**Independent variable re-expression (function)**	**Dependent variable re- expression (function)**
E.3	0.0829	2.6149e-07	BP = 15.59094 df = 1 p-value = 7.9e-05	None	x^2	1/x
E.4.1	0.1398	4.3445e-12	BP = 8.509977 df = 1 p-value = 0.003532	x/1000	None	x^2
E.4.1.2	0.1584	1.0446e-13	BP = 10.46671 df = 1 p-value = 0.001215	None	$X^{1/2}$	x^2
E.4.3	0.0564	1.6139e-05	BP = 9.05031 df = 1 p-value = 0.002627	x/1000	None	x^2
E.5	0.1806	5.5275e-15	BP = 15.765802 df = 1 p-value = 7.2e-05	None	None	x^2
E.6	0.0694	2.6661e-06	BP = 5.716832 df = 1 p-value = 0.016803	None	x^2	1/x
E.8	0.2086	2.5186e-17	BP = 8.331141 df = 1 p-value = 0.003897	x/1000000	x^2	1/x
E.16	0.2021	1.5955e-17	BP = 14.86085 df = 1 p-value = 0.000116	x/1000000	1/x	sqrt(x)
E.16.1	0.1977	3.9432e-17	BP = 14.59149 df = 1 p-value = 0.000134	x/1000	1/x	1/x
E.16.2	0.2242	1.6539e-19	BP = 10.48399 df = 1 p-value = 0.001204	x/1000000	1/x	1/x
E.16.3	0.2182	5.7351e-19	BP = 12.934979 df = 1 p-value = 0.000322	x/1000000	1/x	None

Table 5.5 Model selection. Stepwise procedure for MCMA.

Step	Predictor variable ID	Added/ Removed	R^2	R^2-Adjusted	Cp	AIC	RMSE
1	E_16_2	addition	0.2255	0.2229	99.8668	1983.416	6.5542
2	E_4_1_2	addition	0.2822	0.2773	72.9017	1962.621	6.3204
3	E_2_6	addition	0.3098	0.3028	60.7859	1952.851	6.2081
4	E_3	addition	0.3359	0.3269	49.4527	1943.293	6.1
5	E_16	addition	0.3658	0.355	36.158	1931.462	5.9711
6	E_16_2	removal	0.3658	0.3572	34.1621	1929.466	5.961
7	E_5	addition	0.4173	0.4073	9.8485	1906.068	5.7237
8	E_4_1	addition	0.4244	0.4126	8.2188	1904.39	5.6984
9	E_4_3	addition	0.437	0.4235	3.7771	1899.749	5.6453

The resulting MLR model for ozone in the MCMA is summarized as follows:

$$\ln O_3 = 3.346 \cdot 10^{-16} + 0.082 \cdot {E_{2.3.1}}^2 + 0.2803 \cdot \frac{1}{{E_{2.5.1}}^2} - 0.4998 \cdot \frac{1}{{E_3}^2} + 0.1543 \cdot {E_7}^2 - 0.7351 \cdot \frac{1}{{EMEC_2}^2} + 0.599 \cdot \frac{1}{{EMEC_4}^2 \cdot ENEC_{11.4}} - .2335 \cdot \frac{1}{ENEC_{6.4}} + 0.1981 \cdot \frac{1}{{ENEC_{9.3}}^2} - 0.1144 \cdot \frac{1}{{ENEC_{11.2}}^2} + 0.1724 \quad \textbf{(5.1)}$$

where:

$E_{2.3.1}$	Petroleum products processing (Extra/Pemex Magna b Gasoline)
$E_{2.5.1}$	Petroleum products processing (Pemex Diesel)
E_3	Domestic LP Gas demand
E_7	Domestic Natural Gas Demand (Petroleum Sector)
$EMEC_2$	Total revenues from the supply of goods and services
$EMEC_4$	Average remuneration
$ENEC_{6.4}$	Social benefits excluding worker and administrative employees' wages
$ENEC_{9.3}$	Electrical and telecommunications constructions and auxiliary works
$ENEC_{11.2}$	Water, irrigation and sanitation constructions and auxiliary works
$ENEC_{11.4}$	Transportation constructions and auxiliary works

The Equation (5.1) coefficients of the linear regression model and their corresponding statistics are shown in Tables 5.6 and 5.7.

Table 5.6 Coefficients of the linear regression model for the MCMA.

Variable	Coefficients	Tolerance	VIF	Significance
Intercept	3.346E-16	-	-	-
$E_{2.3.1}$	0.082	0.2579	3.8776	0.4469
$E_{2.5.1}$	0.2803	0.4483	2.2305	0.0008
E_3	-0.4998	0.8021	1.2468	<0.0001
E_7	0.1543	0.4937	2.0255	0.0492
$EMEC_2$	-0.7351	0.2179	4.5883	<0.0001
$EMEC_4$	0.599	0.3882	2.5757	<0.0001
$ENEC_{6.4}$	-0.2335	0.4987	2.0053	0.0031
$ENEC_{9.3}$	0.1981	0.6737	1.4844	0.0035
$ENEC_{11.2}$	-0.1144	0.7698	1.299	0.0684
$ENEC_{11.4}$	0.1724	0.7951	1.2577	0.0057

Table 5.7 Performance statistics for the MCMA O_3 MLR model.

Statistic	Value	Statistic	Value

R^2	0.6395	Global Stat	8.1160
Adjusted R^2	0.6097	p-value	0.0874
F-Test	21.4629	Skewness	2.6713
p-value	<0.0001	p-value	0.1022
Bresuch-Pagan Test	13.5319	Kurtosis	0.0756
p-value	0.1954	p-value	0.7833
Durbin-Watson Test	1.8098	Link Function	2.9331
p-value	0.1061	p-value	0.0868
		Heteroscedasticity	2.4360
		p-value	0.1186

It can be noticed that tests show a notable statistical significance. Figure 5.3 presents QQ plots showed adequate normal residual distribution, as well as a Cook's Distance chart, resulting in values <0.1 for all but 3 observations. The DW statistic (1.81) shows evidence of some correlation, as for a true model with negligible autocorrelation a DW value of 2 would be ideal (Chellakan et al., 2022. Obtained R^2 is spread around 0.6395, which can be considered a good value when compared to the findings of studies using structural equation modeling. (Zhao et al., 2019) obtain an explanation of 42% of the data. Of note, including variables from the Tula Refinery improved the model´s performance and confirms the effect of refinery operations in MCMA's air quality.

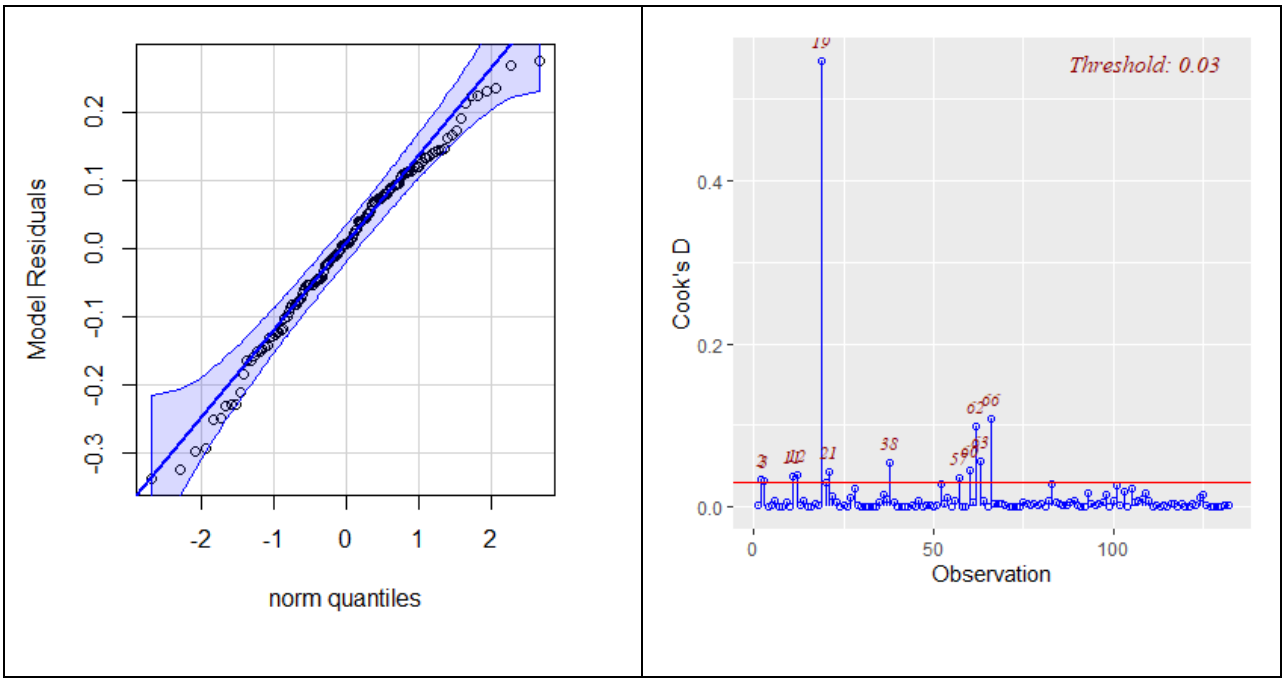

Figure 5.3 Residuals QQ plot and Cook's D Chart for forward-backward stepwise regression for ozone concentration in the MCMA.

While scaling the variables can help computational performance, this might to scaled regression coefficients with magnitudes that do not necessarily equate to the relative importance of the regressor (Montgomery, 2006). In other instances, even the sign of the coefficient might contradict expectations based on experiences. For example, one would expect a proportional relationship between total revenues from the supply of goods and services (EMEC2) and air pollution, but in the MLR model they show a negative relation. This can be due to the existence of regressors not considered in this study due to the lack of available information. However, what can be concluded is that the regressors in the resultant MLR model are important and worth of further analysis to understand sources of pollution, in this case, O_3 concentration.

The robustness of the model was tested through an ensemble model approach, in which 10 randomized subsets that included 70% of all data were create, and a MLR model was created for each data set. The coefficients of the 10 bootstrap regression models were averaged. Each bootstrap model was tested by comparing the remaining

(30%) data points to the model estimates. The resulting average-ensemble model was similarto the original model in terms of coefficient magnitudes, term signs and performance statistics.

5.3.2 MLR model for PM_{10} in the MMA

All 21 possible predictor variables were tested through univariate linear regressions, as described above. Table 5.8 presents descriptive statistics and general information for the MMA selected predictor variables, while Table 5.9 displays model performance statistics, and indicates if any re-expression was performed to increase the regression coefficient or assuring homoscedasticity.

Once the variables with the best results in their univariate linear regressions have been selected, a variety of stepwise forward and backward multivariate linear regression models were built. Each model was tested for statistical performance appraisal by means of its R^2, significance (p-value), F-statistic, AIC, Mallow's Cp, DW, normality of residuals (QQ-plot), presence of influential observations (Cook's) D, homoscedasticity, variables' significance, VIF and tolerance.

For example, a first model using backward stepwise regression method and considering the total of the 21 candidate predictor variables was built seeking to optimize the p-value, and with $\boldsymbol{y'} = \boldsymbol{Ln}(\boldsymbol{y})$. In a second model, forward and backward stepwise regression method was applied to optimize the p-value, and with $\boldsymbol{y'} = \boldsymbol{Ln}(\boldsymbol{y})$. Because this second model included, variables E.2 and E.2.3, the latter being a subset of E.2, the variable E.2 was arbitrarily removed to develop a third model. Also, a mobility variable, E.16.1 was added to produce a more activity-diversified model. Finally, a stepwise forward and backward model was used to produce a linear model $\boldsymbol{y'} = \boldsymbol{y}$.

Table 5.8 Descriptive statistics and general information for the MMA selected predictor variables.

Predictor Variable ID	Variable description	Category	Data source	Data availability (spatial, temporal)	Min	Mean	Max	Standard Deviation
E.1	Process of crude oil and liquids in the Cadereyta refinery (Total). [thousands of barrels per day]	Energy	SIE	Data by State and monthly (1993-2021)	56.69	170.07	239.39	45.7
E.1.2	Process of crude oil and liquids in the Cadereyta refinery (Light). [Thousands of barrels per day]	Energy	SIE	Data by State and monthly (1993-2021)	12.06	83.06	147.8	28.34
E.2	Production of petroleum products by the Cadereyta refinery (Total). [Thousands of barrels of crude oil equivalent per day.]	Energy	SIE	Data by State and monthly (1993-2021)	70.28	174.75	247.31	48.51
E.2.3	Production of petroleum products by the Cadereyta refinery (Gasoline). [Thousands of barrels of crude oil equivalent per day.]	Energy	SIE	Data by State and monthly (1993-2021)	18.77	65.05	97.91	19.51
E.4.2	Volume of internal sales of Petroleum Products by federal entity (Kerosene). [Cubic meters]	Energy	SIE	Data by State and monthly (1993-2021)	0.723	21,421.35	55,053.98	11,327.68
E.4.3	Volume of internal sales of Petroleum Products by federal entity (Diesel). [Cubic meters]	Energy	SIE	Data by State and monthly (1993-2021)	30,689.46	102,741.60	169,691.10	29,686.61
E.4.7	Volume of internal sales of Petroleum Products by federal entity (Asphalts). [Cubic meters]	Energy	SIE	Data by State and monthly (1993-2021)	1,179.40	14,139.74	29,421.76	5,839.12
E.4.10	Volume of internal sales of Petroleum Products by federal entity (Others). [Cubic meters]	Energy	SIE	Data by State and monthly (1993-2021)	4,086.70	70,864.49	229,104.60	26,404.40
E.5	Internal Demand for Natural Gas. [Million cubic feet per day]	Energy	SIE	Data by State and monthly (1993-2021)	426.08	658.73	1037.44	127.1375
E.6	Internal Demand for Natural Gas, Public Electricity Sectors and Electricity Exports. [Million cubic feet per day]	Energy	SIE	Data by State and monthly (1993-2021)	91.68	192.34	317.19	47.15
E.11.1	Hours worked by total employed personnel (manufacturing industries). [Number of hours]	Economy	INEGI	Data by State and monthly (2007-2021)	56,055.50	69,950.72	79,844.31	5,352.08
E.12	Total remuneration depending on the company name (Total).	Economy	INEGI	Data by State and monthly (2007-2021)	5,981,912	9,036,927	17,380,054	2,259,910
E.16.1	Registered motor vehicles in circulation (Cars). [Number of vehicles]	Mobility	INEGI	Data by State and monthly (1991-2021)	612,564	1,290,887	1,917,074	349,716.30
E.27	Electric power users. [Number of users]	Energy	SIE	Data by State and monthly (2002-2017)	1,054,232	1,484,880	1,924,832	250,988.10
E.30	Internal sales of electricity. [Megawatt-hour]	Energy	SIE	Data by State and monthly (2002-2017)	17,500.30	1,301,656	1,900,584	267,445.80
E.31	Total wholesale revenue of trading companies. [Index]	Economy	INEGI	Data by State and monthly (2008-2021)	76.32	102.11	121.34	10.42
E.32	Total retail revenue of trading companies. [Index]	Economy	INEGI	Data by State and monthly (2008-2021)	72.94	107.29	157.79	19.48
E.33	Index of water, irrigation, and sanitation of construction companies. [Index]	Economy	INEGI	Data by State and monthly (2006-2021)	9.49	429.73	1,471.88	333.56
E.35	Transportation and urbanization index of construction companies. [Index]	Economy	INEGI	Data by State and monthly (2006-2021)	40.03	130.3	312.29	57.8
E.39	Public sector index of construction companies. [Index]	Economy	INEGI	Data by State and monthly (2006-2021)	30.45	132.62	305.75	56.99
E.40	Index of total income for temporary accommodation and food and beverage preparation services. [Index]	Economy	INEGI	Data by State and monthly (2013-2021)	29.82	79.05	114.28	18.16

Table 5.9 Univariate model performance statistics for MMA selected predictor variables

Predictor Variable ID	R^2	P value	Homoscedasticity test (Brreusch Pragan test)	Independent variable scaling	Independent variable re-expression (function)	Dependent variable re-expression (function)
E.1	0.278	< 2.2e-16	Pass. BP = 11.439, p-value = 0.0007192	Unscaled.	Without re-expression.	Natural logarithm. $\boldsymbol{y'} = \mathbf{ln}(\boldsymbol{y})$.
E.1.2	0.2619	< 2.2e-16	Pass. BP = 14.358, p-value = 0.0001511	Unscaled.	Without re-expression.	Natural logarithm. $\boldsymbol{y'} = \mathbf{ln}(\boldsymbol{y})$.
E.2	0.3016	< 2.2e-16	Pass. BP = 12.958, p-value = 0.0003185	Unscaled.	Without re-expression.	Natural logarithm. $\boldsymbol{y'} = \mathbf{ln}(\boldsymbol{y})$.
E.2.3	0.2772	< 2.2e-16	Pass. BP = 10.935, p-value = 0.0009436	Unscaled.	Without re-expression.	Natural logarithm. $\boldsymbol{y'} = \mathbf{ln}(\boldsymbol{y})$.
E.4.2	0.0517	0.00180	Pass. BP = 16.757, p-value = 4.249e-05	Scaled x . 1000	Without re-expression.	Natural logarithm. $\boldsymbol{y'} = \mathbf{ln}(\boldsymbol{y})$.
E.4.3	0.0430	0.0007122	Pass. BP = 5.3661, p-value = 0.02053	Scaled x . 1000	Without re-expression.	Natural logarithm. $\boldsymbol{y'} = \mathbf{ln}(\boldsymbol{y})$.
E.4.7	0.1663	1.355e-11	Pass. BP = 10.642, p-value = 0.001105	Scaled x . 1000	Without re-expression.	Natural logarithm. $\boldsymbol{y'} = \mathbf{ln}(\boldsymbol{y})$.
E.4.10	0.1459	4.002e-09	Pass. BP = 7.9499, p-value = 0.004809	Scaled x . 1000	Without re-expression.	Natural logarithm. $\boldsymbol{y'} = \mathbf{ln}(\boldsymbol{y})$.
E.5	0.189	6.631e-13	Pass. BP = 4.4812, p-value = 0.03427	Unscaled.	Without re-expression.	Radical. $\boldsymbol{y'} = \boldsymbol{y^{3/2}}$.
E.6	0.1701	1.215e-11	Pass. BP = 5.1815, p-value = 0.02283	Unscaled.	Without re-expression.	Natural logarithm. $\boldsymbol{y'} = \mathbf{ln}(\boldsymbol{y})$.

E.11.1	0.0806	0.003192	Pass. BP = 9.1214, p-value = 0.002526	Scaled x . 1000	Without re-expression.	Without re-expression.
E.12	0.0437	0.03156	Almost passed. BP = 3.5254, p-value = 0.06044	Scaled x . 1000000	Fraction. x´ = 1 / x	Fraction. y´ = 1 / y
E.16.1	0.1955	4.595e-14	Pass. BP = 5.1678, p-value = 0.02301	Scaled x . 1000	Without re-expression.	Without re-expression.
E.27	0.2976	3.353e-16	Pass. BP = 8.391, p-value = 0.003771	Scaled x . 1000	Without re-expression.	Natural logarithm. $\boldsymbol{y'} = \mathbf{ln}(\boldsymbol{y})$.
E.30	0.3559	< 2.2e-16	Pass. BP = 28.679, p-value = 8.544e-08	Scaled x . 1000	Without re-expression.	Natural logarithm. $\boldsymbol{y'} = \mathbf{ln}(\boldsymbol{y})$.
E.31	0.0782	0.0002408	Pass. BP = 6.2324, p-value = 0.01254	Unscaled.	Without re-expression.	Without re-expression.
E.32	0.1164	6.018e-06	Pass. BP = 6.4255, p-value = 0.01125	Unscaled.	Without re-expression.	Natural logarithm. $\boldsymbol{y'} = \mathbf{ln}(\boldsymbol{y})$.
E.33	0.1307	0.0001208	Pass. BP = 3.9449, p-value = 0.04701	Unscaled.	Without re-expression.	Without re-expression.
E.35	0.2521	1.181e-13	Pass. BP = 5.3651, p-value = 0.02054	Unscaled.	Without re-expression.	Radical. $\boldsymbol{y'} = \boldsymbol{y}^{3/2}$.
E.39	0.2068	3.441e-11	Pass. BP = 3.9988, p-value = 0.04553	Unscaled.	Without re-expression.	Radical. $\boldsymbol{y'} = \boldsymbol{y}^{2}$.

After analyzing models' performance statistics, it was noted that models 1 and 4 clearly underperformed, while models 2 and 3 showed similar performance. However, even when model 2 displayed a slightly better R^2, after comparing variable tolerances VIF, and Cook's D, and overall structure, it was concluded that model 3 provided a better fit. Therefore, the resulting MLR model for PM_{10} in the MMA is presented in Equation (5.2). Tables 5.10 and 5.11 present coefficients of the linear regression model and their corresponding performance statistics.

$$\ln(O_3) = 2.2558 - 0.0333\,E_{16.1} + 0.8133\ln(E_5^{2/3}) - 0.0015\,E_6 - 1.6523\ln(E_{11.1}) + 0.0031\,E_{2.3} + 0.0109\,E_{4.7} + 1.1803\ln(E_{31}) \quad \textbf{(5.2)}$$

where:

E.2	Total petroleum products
E.6	Internal demand of natural gas for electricity production
E.11.1	Hours worked by manufacturing industries personnel
E.5	Total internal demand of natural gas (state)
E.31	Total wholesale revenue of trading companies
E.2.3	Total production of petroleum products
E.4.7	Internal sales of petroleum products (asphalts)

Table 5.10. Coefficients of the linear regression model for the MMA.

Predictor variable ID	Variable coefficient (Estimate)	Std. Error	Tolerance	VIF	Significance
(Intercept)	2.2558	2.4757			0.3648
E.6	-0.0015	0.0004	0.8325	1.2012	0.0022
E.11.1	-1.6523	0.6124	0.1782	5.6129	0.0084
E.5	0.8133	0.3588	0.3448	2.9002	0.0260
E.31	1.1803	0.4949	0.2816	3.5517	0.0193
E.2.3	0.0031	0.0015	0.4847	2.0631	0.0437
E.4.7	0.0109	0.0053	0.9047	1.1054	0.0427
E.16.1	-0.0333	0.2618	0.4100	2.4388	0.8990

Table 5.11 MMA PM_{10} MLR model performance statistics.

R2	0.4440
R^2-Adjusted	0.3982
p-value	8.11e-09
F-statistic	9.696 on 7 and 85 DF
Global Stat	Assumptions acceptable.
Skewness	Assumptions acceptable.
Kurtosis	Assumptions acceptable.
Link Function	Assumptions acceptable.
Heteroscedasticity test	Assumptions acceptable.
Durbin-Watson test	DW = 1.1249, p-value = 3.124e-07

The variables in Equation (5.2) can explain almost half of the variation observed in PM_{10} monthly averaged concentration, which can be considered a good fit. Regarding Table 5.11, the Global Stat assumption indicates if a linear relationship exists between the dependent variable and the independent variables, the Link function assumption verifies that the variable is continuous, and the heteroscedasticity test confirms that the variance of the residuals is constant. Furthermore, the Skewness and Kurtosis assumptions show that the distribution of the residuals is normal, as confirmed by the QQ plot in Figure 5.4. On the other hand, the Durbin-Watson statistic implies that there might be some autocorrelation, however, there are no significant outlier observations in Cook's D chart.

Also, as discussed in section 5.3.1, the resulting model in Equation 5.2 cannot fully explain all intervening pollution processes, and there might be uncertainty in regards to the significance or sign of the regressor coefficients. However, the model identifies that activities of oil production and processing have some relation with O_3 pollution in the AMM, and that this relation should be further explored. Also, variables such as natural gas demand, number of registered vehicles, asphalt sales, wholesale sales index, and hours worked in manufacturing industries can have an impact in pollution levels, too.

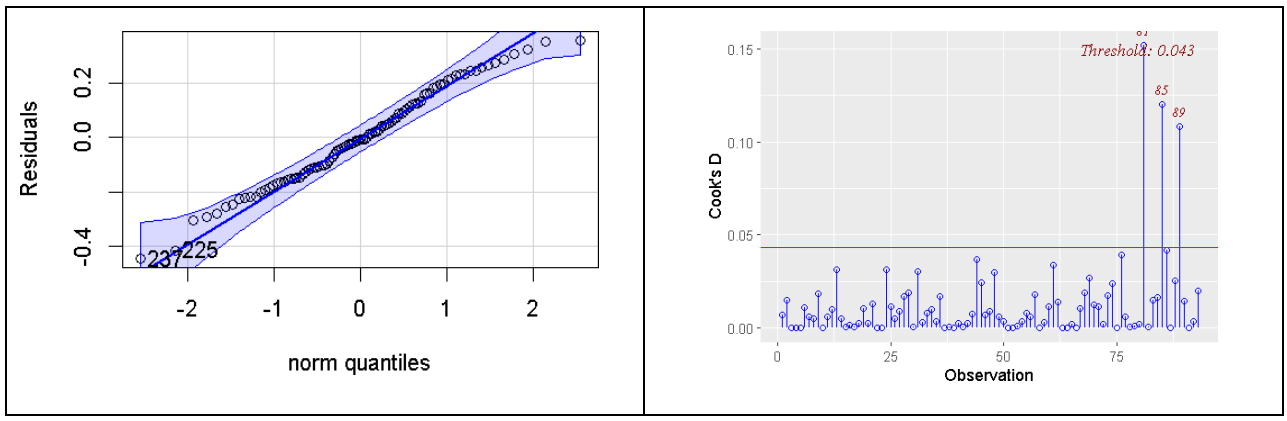

Figure 5.4 Residuals QQ plot and Cook's D Chart for forward-backward stepwise regression for PM_{10} concentration in the MMA.

5.3.3 MLR model for O_3 in the GMA

For the GMA, 17 energy, economic and mobility variables complied with the univariate regression analysis significance criteria to be considered possible predictors in the MLRM, namely homoscedasticity of 0.1 and R^2>0.05. In some cases, re-expressions of the independent and dependent variables were performed. Of note, in the case of GMA, there are no nearby refineries, thus there were no fuel production and processing variables. Table 5.12 presents descriptive statistics and general information for the GMA selected predictor variables, while Table 5.13 displays model performance statistics, and indicates if any re-expression was performed to increase the regression coefficient or assuring homoscedasticity.

Table 5.12 Descriptive statistics and general information for the GMA selected predictor variables.

Predictor Variable ID	Variable description	Category	Data source	Data availability (spatial, temporal)	Min	Mean	Max	Standard Deviation
E.4.1.2	Domestic demand of petroleum products: Pemex Magna (m^3).	Energy	SIE - Instituto Mexicano del Petróleo	Monthly data from 1999 to 2021, selected by state.	2,240	18,100	30,007	5,092
E.4.2.2	Domestic demand of petroleum products: Paraffin oil (m^3).	Energy	SIE - Instituto Mexicano del Petróleo	Monthly data from 2001 to 2021, selected by state.	9,615	32,593	47,009	8,303
E.8	Domestic demand of Natural Gas: Public Sector and Exportations (m^3).	Energy	SIE - Instituto Mexicano del Petróleo	Monthly data from 1996 to 2020, selected by state.	0.1869	1.3239	3.4787	0.865243
E.10	Total employment (Number of workers).	Economy	INEGI - Encuesta Mensual de la Industria Manufacturera	Monthly data from 2013 to 2021, selected by state.	207,259	237,486	266,548	19,451
E.11	Number of hours worked by employed personnel.	Economy	INEGI - Encuesta Mensual	Monthly data from 2007 to 2021, selected	403,73	47,981	55,608	3,926
E.13	Production value of manufactured products (Thousands of nominal pesos).	Economy	INEGI - Encuesta Mensual de la Industria Manufacturera	Monthly data from 2007 to 2021, selected by state.	26,442,469	37,671,564	50,470,161	6279,083
E.16.2	Registered motor vehicles in circulation: Buses	Mobility	INEGI	Monthly data from 1996 to 2021, selected by state.	4,764	10,044	12,616	1,745
E.16.3	Registered motor vehicles in circulation: Loading Vehicles	Mobility	INEGI	Monthly data from 1996 to 2021, selected by state.	363,217	802,096	1,060,772	234,361
E.24	Number of accidents from ground transportation.	Mobility	SCT- Secretaria de Infraestructura, comunicación y transportes	Monthly data from 1997 to 2021, selected by state.	768	3,328	5,430	1,298
E.26	Gross generation of CFE plus PIE's by state (Megawatt-hours.)	Energy	SIE- Electrical National System	Monthly data from 2002 to 2020, selected by state.	16619	85486	591766	118925.8
E.28.2	Irrigation, Water and Sanitation Index.	Economy	INEGI-ENEC	Monthly data from 2006 to 2021, selected by state.	3.42	76.57	230.06	54.96
E.28.3	Electricity and Telecommunications Index.	Economy	INEGI-ENEC	Monthly data from 2006 to 2021, selected by state.	2.95	57.22	180.11	43.33
E.28.5	Oil and Petrochemical Index.	Economy	INEGI-ENEC	Monthly data from 2006 to 2021, selected by state.	0.81	41.80	160.82	51.04
E.28.7	Public Sector Index.	Economy	INEGI-ENEC	Monthly data from 2006 to 2021, selected by state.	19.32	99.36	214.70	47.36
E.28.8	Private Sector Index.	Economy	INEGI-ENEC	Monthly data from 2006 to 2021, selected by state.	51.91	95.81	205.65	23.90
E.30.1	Total income from supply of good and services: Wholesale trade.	Economy	INEGI- EMEC	Monthly data from 2008 to 2021, selected by state.	81.75	109.29	147.14	16.59
E.30.2	Total income from supply of good and services: Retail trade.	Economy	INERGI-EMEC	Monthly data from 2008 to 2021, selected by state.	78.98	107.79	157.87	18.21

Table 5.13 Univariate regression statistics for the GMA selected predictor variables.

Univariate regression model						
Predictor Variable ID	**R^2**	**P value**	**Homocedasticity test (Brreusch Pragan test)**	**Independent variable scaling**	**Independent variable re-expression (function)**	**Dependent variable re-expression (function)**
E.4.1.2	0.002086	0.4556	BP = 20.412, df = 1, p-value = 6.242e-06	1:1000	-	$In(y)$
E.4.2.2	0.07698	7.804e-06	BP = 2.2438, df = 1, p-value = 0.1342	1:1000	-	y^3
E.8	0.01217	0.05757	BP = 2.8327, df = 1, p-value = 0.09236	-	-	$In(y)$
E.10	0.01566	0.1969	BP = 3.2494, df = 1, p-value = 0.07145	1:10000	$In(x)$	$1/y$
E.11	0.02014	0.1429	BP = 3.8071, df = 1, p-value = 0.05103	1:1000	-	$In(y)$
E.13	0.03321	0.05907	BP = 3.8441, df = 1, p-value = 0.04992	1:1000000	-	$In(y)$
E.16.2	0.09941	1.251e-08	BP = 36.116, df = 1, p-value = 1.859e-09	1:100	-	$In(y)$
E.16.3	0.0694	2.38e-06	BP = 25.541, df = 1, p-value = 4.33e-07	1:10000	-	$In(y)$
E.24	0.07775	9.236e-07	BP = 6.9707, df = 1, p-value = 0.008285	1:100	-	$In(y)$
E.26	0.1058	5.137e-07	BP = 2.8372, df = 1, p-value = 0.09211	1:100	$In(x)$	y^2
E.28.2	0.03709	0.007446	BP = 15.692, df = 1, p-value = 7.454e-05	-	$In(x)$	y^3
E.28.3	0.02478	0.0292	BP = 11.822, df = 1, p-value = 0.0005852	-	$In(x)$	y^3
E.28.5	0.02056	0.2296	BP = 4.2134, df = 1, p-value = 0.04011	-	$In(x)$	$1/y$
E.28.7	0.09912	8.683e-06	BP = 7.2438, df = 1, p-value = 0.007115	-	$In(x)$	y^3
E.28.8	0.04435	0.003366	BP = 5.4058, df = 1, p-value = 0.02007	-	$In(x)$	$\sqrt{y}$
E.30.1	0.1874	4.628e-09	BP = 7.1448, df = 1, p-value = 0.007518	-	$In(x)$	y^3
E.30.2	0.1438	3.998e-07	BP = 6.4716, df = 1, p-value = 0.01096	-	$In(x)$	y^3

As in the case of MCMA and MMA, stepwise forward and backward regression models were computed. Considering variables listed in Table 5.13 as possible predictor variables, a recommended model for monthly-averaged O_3 concentration in the GMA is:

$$[\boldsymbol{O_3}] = \mathbf{239.636} + \mathbf{0.372}\,(\boldsymbol{E_{30.1}}) - \mathbf{0.640}\,(\boldsymbol{E_{28.8}}) - \mathbf{0.151}(\boldsymbol{E_{28.7}}) - \mathbf{0.294}\,(\boldsymbol{E_{28.3}}) + \mathbf{0.113}\,(\boldsymbol{E_{28.5}}) + \mathbf{62.073}(\boldsymbol{E_{10}}) + \boldsymbol{E_{4.1.2}} + \boldsymbol{E_{13}} + \boldsymbol{E_{24}} + \mathbf{0.509}\,(\boldsymbol{E_{30.2}}) \quad \textbf{(5.3)}$$

where:

E.28.3	Electricity and telecommunications index.
E.12	Total remuneration depending on denomination
E.16.3	Registered motor vehicles in circulation: Loading vehicles
E.13	Production value of manufactured products
E.30.1	Total income from supply of good and services: Wholesale trade.
E.16.2	Registered motor vehicles in circulation: Buses
E.4.2.	Domestic demand of petroleum products: other kerosenes
E.28.1	Building index
E.28.6	Other constructions Index
E.28.2	Irrigation, water and sanitation Index.
E.16.1	Registered motor vehicles in circulation: Cars
E.30.2	Total income from supply of good and services: Retail trade.

This model in Equation (5.3) was obtained through a forward stepwise regression, as detailed in Table 5.14 below. The linear regression coefficients and its corresponding statistics are shown in Table 5.15, while the model's performance statistics are shown in Table 5.16. As in the cases of MCMA and MMA, an ensemble model was built to test for robustness.

Table 5.14 Forward stepwise regression for MLR O_3 model in the GMA.

Run	Predictor variable ID	Added/Removed	R2	R2-Adjusted	Cp	AIC	RMSE
1	E.30.1	Added	0.1437	0.1385	98.9221	1028.3974	5.1034
2	E.28.8	Added	0.2249	0.2155	75.9754	1013.6493	4.8699
3	E.28.7	Added	0.2521	0.2385	69.6176	1009.6430	4.7982
4	E.28.3	Added	0.2942	0.2769	58.7106	1001.9238	4.6757
5	E.28.5	Added	0.2419	0.1630	23.9448	329.2325	4.7530
6	E.10	Added	0.3073	0.1589	13.0013	212.3100	4.4685
7	E.4.1.2	Added	0.3845	0.2249	11.2115	210.1738	4.2894
8	E.13	Added	0.4694	0.3061	9.0460	206.9813	4.0586
9	E.24	Added	0.5320	0.3635	7.9713	204.5846	3.8870
10	E.30.2	Added	0.5576	0.3732	8.7180	204.6210	3.8575

Table 5.15. Coefficients of the linear regression model for the GMA

Predictor variable ID	Variable coefficient	Tolerance	VIF	Significancia
E.30.1	0.372	0.09089516	11.001686	0.494
E.28.8	-0.640	0.82932225	1.205804	0.942
E.28.7	-0.151	0.32584571	3.068937	0.258
E.28.3	-0.294	0.31861664	3.138568	0.017
E.28.5	0.113	0.48666918	2.054784	0.854
E.10	62.073	0.12367405	8.085770	0.031
E.4.1.2	0.000	0.90693557	1.102614	0.025
E.13	0.000	0.55720345	1.794677	0.084
E.24	0.000	0.80370397	1.244239	0.071

Table 5.16. GMA PM_{10} MLR model performance statistics.

Statistic	Value	p-value
Global Stat	1.04E+01	0.03385
Skewness	1.56E+00	0.21184
Kurtosis	7.54E-04	0.9781
Link function	6.48E+00	0.01093
Heteroscedasticity	2.39E+00	0.12221
R2	0.6142	
R2 Adjusted	0.5563	
F-statistic	10.61	3.30E-12
Breusch-Pagan	20	6.71E-02
Durbin-Watson	1.4194	1.16E-04
Mallows Cp	16.0085	
RMSE	0.109	

As can be noticed, consumption of fuels (primarily, internal demand of natural gas and gasoline display the most significant correlation with O_3 pollution. Of note, in the resulting model, some economic activity variables correlated negatively with air pollution. Again, this can be to some lack of information for all relevant regressors not considered in this study due to the lack of available information. However, what can be concluded is that the regressors in the resultant MLR model are important and worth of further analysis to understand sources of pollution.

Finally, Table 5.17 presents the summary of model performance results for MCMA, MMA and GMA models.

Table 5.17 Summary of model performance statistics for MCMA, MMA and GMA MLR models.

Model statistics	**MCMA Forward-Backward Model**	**MMA Forward-Backward Model**	**GMA Forward model**
R^2	0.4402	0.4440	0.6142
R^2 Adjusted	0.4268	0.3982	0.5563
F-statistic	32.8038	9.696	10.61
Durbin-Watson	1.0978	1.1249	1.4194

5.4 Conclusions and discussion

Air quality in metropolitan areas is the product of multiple confluence factors, both natural and anthropogenic. This study attempted to identify economic and energy activities with statistically significance in air quality, in three Mexican Metropolitan areas: MCMA, MMA and GMA. Even if sources are not explicitly identified, socioeconomic indicators among air quality and economic activity, as they change through time, allow for the study of air pollution responses to forcings.

The resulting MLRM models compare favorably to other studies found in literature, e.g. Zhao et al. 2012, in terms of R^2 and p-value significance. It is concluded that, although the three areas have distinct geographical locations and characteristic, common air pollution drivers appear: fuel production and processing resulted an important factor in areas nearby refinery installations, as well as fuel use (e.g. gasoline and natural gas demand). It is also demonstrated that the use of energy and economy information, routinely recorded by government agencies can provide further insights on the causes of air pollution, and that these economic and energy activities alone can explain a substantial fraction of air pollution. Identifying and quantifying the causes of urban air pollution and allows the formulation of pollution control measures in Mexico.

For example, in the case of the MMA, although there are experimental studies that suggest an important contribution of the Cadereyta Refinery in MMA air quality, there are no deterministic simulations that can further support this hypothesis. The MMA-MLR model states that there are some variables associated to the Cadereyta activity that show that is a significant activity, which provides evidence to the need to develop public policies and measures.

.5.5 References

Benítez-García, S. E., Kanda, I., Wakamatsu, S., Okazaki, Y., & Kawano, M. (2014). Analysis of criteria air pollutant trends in three Mexican metropolitan areas. Atmosphere, 5(4), 806-829.

Cole, M. A., & Neumayer, E. (2004). Examining the impact of demographic factors on air pollution. *Population and Environment*, 26(1), 5-21.

Cramer, J. C. (1998). Population growth and air quality in California. *Demography*, *35*(1), 45-56.

Hernández Paniagua, I. Y., Clemitshaw, K. C., & Mendoza, A. (2017). Observed trends in ground-level O 3 in Monterrey, Mexico, during 1993–2014: comparison with Mexico City and Guadalajara. *Atmospheric Chemistry and Physics*, *17*(14), 9163-9185.

Herrera, P., Rojo, J., & Scapini, V. (2022). Relationship between pollution levels and poverty: regions of Antofagasta, Valparaiso and Biobio, Chile. *International Journal of Energy Production and Management. 2022. Vol. 7. Iss. 2*, *7*(2), 176-184.

Kadaverugu, R., Sharma, A., Matli, C., & Biniwale, R. (2019). High resolution urban air quality modeling by coupling CFD and mesoscale models: A review. Asia-Pacific Journal of Atmospheric Sciences, 55(4), 539-556

Mancilla, Y., Hernandez Paniagua, I. Y., & Mendoza, A. (2019). Spatial differences in ambient coarse and fine particles in the Monterrey metropolitan area, Mexico: Implications for source contribution. *Journal of the Air & Waste Management Association*, *69*(5), 548-564.

Montgomery, D. C. (2006). *Introduction to linear regression analysis* (4th ed.). John Wiley & Sons, Inc.

Murillo-Gómez, E., Palomar-Ramírez, M., & Ramos-Flores, M. (2022). Assessment on the Distribution and Accessibility to Green Spaces in Mexico's Most Populated Metropolitan Zones. In International Conference on Geospatial Information Sciences (pp. 3-14). Springer, Cham.

Ortega Montoya, C. Y., López-Pérez, A. O., Ugalde Monzalvo, M., & Ruvalcaba Sánchez, M. L. G. (2021). Multidimensional Urban Exposure Analysis of

Industrial Chemical Risk Scenarios in Mexican Metropolitan Areas. *International Journal of Environmental Research and Public Health*, *18*(11), 5674.

Ramírez Sáiz, J. M., & Safa Barraza, P. (2011). Realidades y retos de las áreas metropolitanas: ciudad de México, Guadalajara y Monterrey. *Desacatos*, (36), 131-148.

Shpak, N., Ohinok, S., Kulyniak, I., Sroka, W., Fedun, Y., Ginevičius, R., & Cygler, J. (2022). CO2 Emissions and Macroeconomic Indicators: Analysis of the Most Polluted Regions in the World. *Energies*, *15*(8), 2928.

Stolz, T., Huertas, M. E., & Mendoza, A. (2020). Assessment of air quality monitoring networks using an ensemble clustering method in the three major metropolitan areas of Mexico. *Atmospheric Pollution Research*, *11*(8), 1271-1280.

Vega, E., Ramírez, O., Sánchez-Reyna, G., Chow, J. C., Watson, J. G., López-Veneroni, D., & Jaimes-Palomera, M. (2022). Volatile Organic Compounds and Carbonyls Pollution in Mexico City and an Urban Industrialized Area of Central Mexico. Aerosol and Air Quality Research, 22, 210386.

WHO Global Health Observatory. Air pollution data portal. Geneva: World Health Organization, 2022 (https://www.who.int/data/gho/data/themes/air-pollution, accessed 5 November 2022).

World Health Organization. (2022). World health statistics 2022: monitoring health for the SDGs, sustainable development goals.

Yang, M., Wang, W., Li, Y., Du, Y., & Tian, F. (2022). Revealing the Impact of Socio-Economic Metrics on the Air Quality on Northeast China Using Multivariate Statistical Analysis. *Polish Journal of Environmental Studies*, *31*(4).

Zhou, M., Li, Y., & Zhang, F. (2022). Spatiotemporal Variation in Ground Level Ozone and Its Driving Factors: A Comparative Study of Coastal and Inland Cities in Eastern China. *International journal of environmental research and public health*, *19*(15), 9687.

This page intentionally left blank

6. Conclusions and future work

6.1. Conclusions

Deterministic air quality models have multiple applications: they can be used to recreate specific episodes, validate emission inventories, and test emission changes scenarios. Detailed, spatially and temporally-resolved emission inventories are a key input to deterministic AQM, however uncertainties in inventories remains a current issue, as exemplified by experiences documented in literature, This is especially relevant for emission inventories developed in the MMA and the State of Nuevo Leon during years 2005 to 2019. For example, emission inventories with base year 2005, presented differences up to +600% in the case of gaseous mobile emissions, among other discrepancies. More recent emission inventories prepared by different government agencies still showed differences ranging from -3.6% to +51.7% for total emissions of specific pollutants between different inventories.

Inverse modeling can be applied with satisfactory results to improve emissions estimates. As study case, some regularization methods (Tikhonov regularization, Truncated Singular Value Decomposition, and Damped Singular Value Decomposition) in combination with regularization parameter selection methods (Generalized Cross Validation, L-Curve, and Normalized Cumulative Periodograms), along a Bounded Variable Least Squares (BVLS) method, were used with a deterministic photochemical air quality model to compute scaling factors for the improvement of a criteria-pollutant emission inventory for Guadalajara Metropolitan Area, Mexico.

The inverse modeling with regularization approach was able to adequately resolve ozone concentrations, a secondary pollutant, by adjusting its precursor emissions, obtaining Daily Indices of Agreement up to 0.95 (compared to 0.89 of the base case). Also, the non-systematic error was reduced. The experiment also showed that the BVLS consistently showed the best agreement among the other mathematical techniques tested, and that regularization methods demonstrated

almost indistinct behavior patterns among them. Nonetheless, the choice of the regularization parameter was found to explain most of the variance shown among the different tested schemes, with the techniques using the LC method exhibiting better agreement between the observed and simulated values than their NCP and GCV counterparts.

However, this research also shows that mathematical techniques, such as regularization methods, cannot fully resolve all inconsistencies, such as uncertainties in specific emission processes. Therefore, alternative approaches, such as observational procedures, are needed. Recent advances in data science allows for the use of supervised machine learning methods, such as Multivariate Linear Regression Model (MLRM) along socioeconomic historical data to explore the evolution of pollution sources through time. Actually, both -inverse modeling with regularization and MLR- approaches are driven by observations. In the case of inverse modeling, observations are used to adjust emissions. In the case of MLRM, observations are used to train the model.

Here, MLR models were built by correlating socioeconomic and energy monthly data with monthly-averaged pollutant concentrations in MCMA, MMA, and GMA. Ensemble models showed that the obtained models were robust. The resulting models included energy variables associated to fuel production (when a nearby refinery existed), mobility, and economic activity. The resulting MLRM models compare favorably to other studies found in literature, (e.g. Zhao et al. 2012, Chellakan 2022) in terms of R^2 and p-value significance. The statistical performance of MCMA, MMA and GMA were similar among the three geographical areas.

Deterministic AQM and supervised ML methods, as MLR models, are two different tools, that can be complementary as they try to compensate the other's shortcomings, but together can be used to create relevant public policies.

6.2 Future work

This research demonstrated that inverse modeling can be applied with satisfactory results to improve emissions inventories, one of the major sources of uncertainty in air quality modeling applications.

A follow up to this work is the development of structural equation models (SEM). SEM seek to explain theoretically defined cause-effect relationships between latent variables (LV) through multivariate statistical modeling techniques. The LV of a system explain the characteristics of a system but are characterized by not being directly quantifiable. Within a system, LVs can be quantified, indirectly, by means of observable and measurable variables known as manifest variables (MVs), which are sometimes affected by other LV (Jiao et al, 2016). SEM has the advantage over conventional multivariable regression models in that direct and indirect relationships that influence between all the factors analyzed can be traced, while conventional regression only quantifies direct relationships.

The SEM methodology is currently used in various areas of science, particularly in the human behavioral sciences (Huijts et al., 2014), but is also being explored to analyze social, economic, and urban factors, among others, and their possible direct and indirect contributions to air quality (Shi et al.., 2019; Zhao et al., 2018). However, the applications vary as different regions have different indicators, with different resolution (spatial and temporal), which requires an evaluation of the type of information and type of models most appropriate to represent the local contexts.

Another natural branch of further research is the use of land-use regression (LUR) models. LUR models can combine air pollution monitoring data, from typically 20 to 100 sites, along stochastic models using predictor variables, such as traffic representations, population density, land use, physical geography and climate or meteorological variables, usually obtained through geographic information systems (GIS). LUR have been successfully used to model criteria pollutant concentrations in different geographic areas (e.g. Hinojosa-Baliño et al., 2019; Wong et al. 2021, Zhang et al. 2018).

On a last note, it cannot be understated the evident need to continue building capacities in the topics of development and validation of emission inventories and air quality modeling. According to a study from INECC-PNUD (2018), in 2017, there were only three research institutions with sustained work in air quality modeling applications in Mexico: Centro de Ciencias de la Atmósfera of Universidad Nacional Autónoma de México, Instituto Nacional del Petróleo, and Tecnológico de Monterrey, Campus Monterrey. Moreover, only 4 out of existing 20 ProAires at the time had reported the use of air quality models for the evaluation of emission reduction scenarios and for routine air quality management.

In order to advance air quality modeling capabilities, attention to infrastructure development, training of human resources, continuity of existing research groups, and formation of new research groups in academic and government institutions, must be given.

6.3 References

Hinojosa-Baliño, I., Infante-Vázquez, O., & Vallejo, M. (2019). Distribution of PM2. 5 air pollution in Mexico City: Spatial analysis with land-use regression model. *Applied sciences*, *9*(14), 2936.

Huijts, N. M., Molin, E. J., & van Wee, B. (2014). Hydrogen fuel station acceptance: A structural equation model based on the technology acceptance framework. *Journal of Environmental Psychology*, *38*, 153-166.

INECC-PNUD México (2018) Diagnóstico del estado del arte de la química atmosférica en México con relación a los gases de efecto invernadero y los contaminantes climáticos. Proyecto 85488 “Sexta Comunicación Nacional de México ante la Convención Marco de las Naciones Unidas sobre el Cambio Climático”, Luis Gerardo Ruiz Suárez. México.

Jiao, L., Shen, L., Shuai, C., & He, B. (2016). A novel approach for assessing the performance of sustainable urbanization based on structural equation modeling: A China case study. *Sustainability*, *8*(9), 910.

Shi, T., Liu, M., Hu, Y., Li, C., Zhang, C., & Ren, B. (2019). Spatiotemporal pattern of fine particulate matter and impact of urban socioeconomic factors in China. *International Journal of Environmental Research and Public Health*, *16*(7), 1099.

Wong, P. Y., Lee, H. Y., Chen, Y. C., Zeng, Y. T., Chern, Y. R., Chen, N. T., ... & Wu, C. D. (2021). Using a land use regression model with machine learning to estimate ground level PM2. 5. *Environmental Pollution*, *277*, 116846.

Zhang, Z., Wang, J., Hart, J. E., Laden, F., Zhao, C., Li, T., ... & Chen, K. (2018). National scale spatiotemporal land-use regression model for PM2. 5, PM10 and NO2 concentration in China. *Atmospheric Environment*, *192*, 48-54.

Zhao, S., Liu, S., Hou, X., Cheng, F., Wu, X., Dong, S., & Beazley, R. (2018). Temporal dynamics of SO2 and NOX pollution and contributions of driving forces in urban areas in China. *Environmental pollution*, *242*, 239-248.

This page intentionally left blank

Publications

Published papers derived from this dissertation

Vanoye, A.Y., Cárdenas, S.S., González, D.G., Xicoténcatl, C.I., Mendoza, A. (2023) Impacts of economic and sociodemographic variables on air quality in 3 Mexican cities: Monterrey, Guadalajara, and Mexico City. In preparation.

Vanoye, A.Y., & Mendoza, A. (2014). Application of direct regularization techniques and bounded–variable least squares for inverse modeling of an urban emissions inventory. *Atmospheric Pollution Research*, *5*(2), 219-225.

Sierra, A., **Vanoye, A.Y.,** & Mendoza, A. (2013). Ozone sensitivity to its precursor emissions in northeastern Mexico for a summer air pollution episode. *Journal of the Air & Waste Management Association*, *63*(10), 1221-1233.

Vanoye, A.Y., & Mendoza, A. (2010) Revisión de Métodos de Regularización Directa y sus Aplicaciones en las Ciencias Atmosféricas. Extended abstract presented in L Convención Anual IMIQ. Monterrey, México.

Other contributions

Krstikj, A., Sosa Godina, J., García Bañuelos, L., González Peña, O. I., Quintero Milián, H. N., Urbina Coronado, P. D., & **Vanoye García, A.Y.** (2022). Analysis of Competency Assessment of Educational Innovation in Upper Secondary School and Higher Education: A Mapping Review. *Sustainability*, *14*(13), 8089.

Carmona, J. M., Gupta, P., Lozano-García, D. F., **Vanoye, A.Y.,** Hernández-Paniagua, I. Y., & Mendoza, A. (2021). Evaluation of MODIS aerosol optical depth and surface data using an ensemble modeling approach to assess PM2. 5 temporal and spatial distributions. *Remote Sensing*, *13*(16), 3102.

Carmona, J. M., Gupta, P., Lozano-García, D. F., **Vanoye, A.Y**., Yépez, F. D., & Mendoza, A. (2020). Spatial and temporal distribution of PM2.5 pollution over northeastern Mexico: Application of MERRA-2 reanalysis datasets. *Remote Sensing*, *12*(14), 2286.

Carmona, J. M., **Vanoye, A.Y**., Lozano, F., & Mendoza, A. (2015). Dust emission modeling for the western border region of Mexico and the USA. *Environmental Earth Sciences*, *74*(2), 1687-1697.

Vanoye, A.Y., & Mendoza, A. (2009). Mesoscale meteorological simulations of summer ozone episodes in Mexicali and Monterrey, Mexico: analysis of model sensitivity to grid resolution and parameterization schemes. *Water, Air, & Soil Pollution: Focus*, *9*(3), 185-202.

Book chapters

Mendoza, A., Chandru, S., Hu, Y., **Vanoye, A.Y**., & Russell, A. G. (2011). Modeling the dynamics of air pollutants: Trans-boundary impacts in the Mexicali-Imperial valley border region. In *Advanced Air Pollution*. IntechOpen.

Conference Proceedings

Álvarez, A., Webb, C.A., **Vanoye, A.Y**. (2023) Simulaciones interactivas como recurso para la transformación digital de la enseñanza en ingeniería. *Memorias del IX Congreso Internacional de Innovación Educativa.* Monterrey, México. Monterrey, México.

Carmona, J.M., **Vanoye, A.Y**., Yépez, F.D. (2022) Contaminación atmosférica de Monterrey y su relación con los incendios forestales de 2021-2022. *XX Simposio Internacional SELPER*. Monterrey, México.

Domínguez, G., **Vanoye, A.Y**., Carrera, H.E., Niño, E. (2021). Usando inteligencia artificial para analizar las expresiones faciales de los estudiantes y mejorar

su atención en clases. *Memorias del VIII Congreso Internacional de Innovación Educativa*. Monterrey, México.

Clarke, E., Nieves, D.A., Cervantes, P.A., Cuevas, M.L.O., **Vanoye, A.Y.** (2021) Learning process of causes, consequences and solutions to climate change of undergraduate students without background in the subject. *IEEE Global Engineering Education Conference* (EDUCON). pp. 1035-1039.

Vanoye, A.Y., Carrera, H.E., Romero, M.A., Torres, G. (2020) Haciendo biodiésel con realidad virtual. *Memorias del VII Congreso Internacional de Innovación Educativa*. Monterrey, México

Carmona, J. M., Lozano, D. F., L., **Vanoye, A.Y**., Yépez, F.D., Mendoza, A. (2019) Mapeo de Concentraciones y Determinación de Variables Influyentes en la Distribución Regional de PM2.5 utilizando Redes Neuronales y el Modelo de Re-análisis MERRA-2. *2019 Congreso Colombiano y Conferencia Internacional de Calidad de Aire y Salud Pública (CASP)*. IEEE.

Delgado, M., **Vanoye, A.Y**., Jarquín, I. (2019) Enseñanza colaborativa en un esquema internacional enriquecida con tecnología. Memorias del VI Congreso Internacional de Innovación Educativa. Monterrey, México.

Alvarez, J.A., **Vanoye, A.Y**., Webb, C.A. (2019) Aprendizaje colaborativo en educación a distancia con dispositivos gráficos y tecnología de video inmersivo. *Memorias VI Congreso Internacional de Innovación Educativa*. México.

Domínguez, G., **Vanoye, A.Y**., Carrera, H.E., Niño, E. (2019) Primeros pasos en la implementación de tecnologías de reconocimiento facial e inteligencia artificial para medir atención y aprendizaje en clase. *Memorias del VI Congreso Internacional de Innovación Educativa*. Monterrey, México.

Vanoye, A.Y., Carrera, H.E., Romero, M.A. (2018) Experiencia preliminar del uso de realidad virtual en ingeniería química. *Memorias del V Congreso Internacional de Innovación Educativa*. Monterrey, México.

Vanoye, A.Y. (2016) Sostenibilidad en Práctica: One World Challenge. *Memorias del III Congreso Internacional de Innovación Educativa*. México, D.F.

Vanoye, A.Y., Mendoza, A. (2008) MM5 Simulations of Summer Ozone Episodes in Mexicali and Monterrey, México. *XVII Congreso Mexicano de Mete*orología. Monterrey, México.

Vanoye, A.Y., Mendoza, A. (2008) Modeling of windblown dust events in the Nogales Border Region. *Proceedings of the Fourth International Conference on Environmental Science and Technology*. Houston, Texas.

Vanoye, A.Y., Mendoza, A. (2008) Meteorological Patterns During High Ozone Concentration Episodes in Mexicali and Monterrey, Mexico. *Proceedings of the Fourth International Conference on Environmental Science and Technology*. Houston, Texas.

Lozano García, D.F.; Abad, N.; Hori, M.C.; Marvin, S.; **Vanoye, A.Y.;** Mendoza, A. (2008) Nuevos criterios para el diagnóstico ecológico basados en SIG's: fragmentación, balance hídrico y cuencas atmosféricas. Compendio de Resúmenes del 38° Congreso de Investigación y Desarrollo del Tecnológico de Monterrey. Instituto Tecnológico y de Estudios Superiores de Monterrey. Monterrey, N.L, p. 266.

Opinion articles

Vanoye, A.Y. (2018) Enseñanza interdisciplinaria sobre la Tierra para un futuro sustentable. *EduBits Observatorio*. Instituto para el Futuro de la Educación.

Vanoye, A.Y. (2017) El desperdicio de alimentos y la responsabilidad corporativa. *Newsweek en Español*. Suplemento Especial ESR.

Vanoye, A.Y., Martínez, J., Martínez, M.A., Mendoza, A. (2008) Análisis meteorológico del ventarrón ocurrido el 18 de marzo de 2008 en Monterrey, México. *Calidad Ambiental*, 14 (6), 16-20.

Vita

Ana Yael Vanoye García was born in Nuevo Laredo, Tamaulipas, México, on December 1, 1982. She earned a Chemical Engineering degree from the Instituto Tecnológico y de Estudios Superiores de Monterrey, Monterrey Campus in 2004, and a Master in Sciences with major in Environmental Systems in 2007, also from Instituto Tecnológico y de Estudios Superiores de Monterrey, Campus Monterrey. She has been a visiting scholar at University of Texas at Austin, Georgia Institute of Technology, and University of California Irvine.

Since 2011 to date, she is an adjunct professor in the Departments of Sustainable Technologies & Civil Engineering and Sciences of Tecnológico de Monterrey, teaching climate change, sustainability, physics, and chemical engineering courses. Over the course of her academic career, she has advised 40+ Chemical Engineering and Sustainable Development undergraduate capstone projects, tutored community-service student projects, and contributed in the design of new courses and special activities. She received the School of Engineering & Sciences' Intellectual Vitality, Service and Leadership Recognition for outstanding work during semesters August-December 2016, August-December 2020, and February-June 2021, and coauthored a QS Reimagine Education Awards 2020 shortlisted project.

Ana Yael has also worked as private environmental consultant for industry, local government, and NGOs. Recently, she served as Executive Coordinator of Premio Nacional Juvenil del Agua 2020 and 2021, and as Climate Change Coordinator for the city of San Pedro Garza García, Nuevo León. Her latest appointment is as Subject Matter Expert for educational innovation management platforms at Tecnológico de Monterrey.

Her permanent contact is: anayael@gmail.com

This document was typed in using Microsoft Word by Ana Yael Vanoye García.