Landsat-8 ©NASA

Remote Sensing applied to the study of environment-sensitive chronic diseases: A case study applied to Quito, Ecuador

Cesar Ivan Alvarez Mendoza

Programa Doutoral em Engenharia Geográfica

Departamento de Geociências, Ambiente e Ordenamento do Território

2019

Supervisor

Ana Claudia Teodoro, Professor Auxiliar

Department of Geosciences, Environment and Spatial Planning

University of Porto, Faculty of Sciences, Porto, Portugal

Co-supervisor

Alfonso Tierra, Professor Principal

Departamento de Ciencias de la Tierra y de la Construcción,

Universidad de las Fuerzas Armadas ESPE, Sangolquí, Ecuador

U.PORTO

FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

REMOTE SENSING APPLIED TO THE STUDY OF ENVIRONMENT-SENSITIVE CHRONIC DISEASES: A CASE STUDY APPLIED TO QUITO, ECUADOR

PhD Thesis submitted to

University of Porto

Faculty of Sciences

Doctoral Programme in

Surveying Engineering

César Iván Alvarez Mendoza

July 2019

Department of Geosciences, Environment and Spatial Planning

Supervisor
Ana Claudia Teodoro, Professor Auxiliar com Agregação
Department of Geosciences, Environment and Spatial Planning
University of Porto, Faculty of Sciences, Porto, Portugal

Co-supervisor
Alfonso Tierra, Professor Principal
Departamento de Ciencias de la Tierra y de la Construcción,
Universidad de las Fuerzas Armadas ESPE, Sangolquí, Ecuador

July, 2019

Acknowledgements

I would like to thank to my supervisor Professor Ana Cláudia Teodoro for encourage and motivate me to develop this work, for her dedication, patience and professionalism and to my co-supervisor in Ecuador Professor Alfonso Tierra for the help provided in the bachelor degree and the support given during the development of this thesis.

A special thanks to Professor Alberto Freitas and Professor Joao Fonseca from the Medical School, University of Porto, for the contributions and help in the development of the mathematical models and health data provision, respectively.

I would like to thank to all the people who helped me during the development of the work included in this thesis and to my Portuguese friends who helped me to be as at home during my stay in Portugal.

A great thanks to Universidad Politecnica Salesiana for the consideration to obtain a scholarship to the studies in Portugal.

A very special word of grateful to God and to my parents; Cesar and Lidu for supporting me in my life and to my brother Byron for the company in this walk until here.

Finally, but the more importantly, I want to dedicate this work and the most special words of grateful to my family for their support, time, love, comprehension. Specially, my words are to my wife Pauly and my little kids Naty and Feli. Without them, I could not do this achievement real. They are the best of my life, my love and happiness. Thanks for all my loves.

Resumo

Dados de Deteção Remota (DR) têm sido utilizados frequentemente para estudos epidemiológicos, particularmente na avaliação da relação entre doenças infecciosas e o meio ambiente. No entanto, a sua aplicação é ainda limitada a variáveis pré-determinadas/processadas, como por exemplo, índices de vegetação. O principal objetivo deste projeto foi avaliar a aplicabilidade dos dados de DR (apropriadamente calibrados e processados para condições locais em Quito, Equador) no estudo de doenças respiratórias crónicas sensíveis ao ambiente (asma e bronquite como as principais). Para isso, uma revisão aprofundada da literatura para estudar quais os dados de DR e os algoritmos usados para estimar várias variáveis ambientais relacionadas com doenças prevalentes (por exemplo, O_3, PM) foi realizada. Com recurso a bases de dados de saúde (por exemplo, a alta hospitalar), diferentes modelos foram implementados e testados. Além disso, vários algoritmos de *machine learning*, tais como *multiple linear regression*, *partial least square*, *artificial neural network, logistic regression, support vector regression* e *random forest*, foram implementados com o objetivo de encontrar os modelos mais adequados. O modelo final escolhido (*suport vector regression*) permite obter o mapeamento espacial das doenças respiratórias crónicas entre 2013 e 2017 em Quito, Equador. Este trabalho apresenta assim um novo conceito no uso de dados de RS em aplicações ao ambiente e à medicina, e na proposta de diferentes relações com variáveis ambientais.

Palavras Chave: Deteção Remota por satélite, poluição do ar, *machine learning*, análise espacial, doenças respiratória

Abstract

Remote Sensing (RS) data have been frequently used in epidemiological studies, specifically in the assessment of the relationship between infectious disease and the environment. However, their application is limited to pre-determined/processed variables, as vegetation indexes. The main objective of this work was to evaluate the applicability of RS data (appropriately calibrated and processed for local conditions in Quito, Ecuador) in the study of environment-sensitive chronic respiratory diseases (asthma and bronchitis). For this, a comprehensive review of the RS data and the algorithms available used to retrieve several environment variables related to prevalent diseases (O_3, PM), were performed. Using a health database (hospital discharge), different models were computed and tested. Several machine learning methods, as multiple linear regression, partial least squares, artificial neural network, logistic regression, support vector regression and random forest, were applied to find the most adequate models. The final model (support vector regression) allowed to obtain a spatial mapping of the chronic respiratory diseases between 2013 to 2017, in Quito, Ecuador. This work presents a new concept in the use of RS data in different fields like environment and health and in the proposal of different relationships considering environmental variables.

Keywords: Satellite remote sensing, air pollution, machine learning, spatial analysis, respiratory diseases

Contents

List of figures

List of tables

List of acronyms

AIC	Akaike Information Criterion
ANN	Artificial Neural Network
AOT	Aerosol Optical Thickness
AQMN	Air Quality Monitoring Network
B9	Cirrus band Landsat-8
BIC	Bayesian Information Criterion
CO	Carbon monoxide
CO_2	Carbon dioxide
CRD	Chronic Respiratory Disease
ETM+	Enhanced Thematic Mapper Plus
EVI	Enhanced Vegetation Index
LST	Land Surface Temperature
LUR	Land Use Regression
MLP	MultiLayer Perceptron
MLR	Multiple Linear Regression
MODIS	Moderate Resolution Imaging Spectroradiometer
NDVI	Normalized Difference Vegetation Index
NIR	Near-InfraRed
NO_2	Nitrogen dioxide
O_3	Ozone
OLI	Operational Land Imager
PLS	Partial Least Square
PM10	Particulate Matter 10 micrometers or less
PM2.5	Particulate Matter 2.5 micrometers or less
R^2	Coefficient of Determination – R squared
RFR	Random Forest Regression
RMSE	Root-Mean-Square Error
RS	Satellite Remote Sensing
SAVI	Soil-Adjusted Vegetation Index
SO_2	Sulfur dioxide
STW	Stepwise regression
SVR	Support Vector Regression
SWIR	Short-Wave InfraRed
TIRS	Thermal Infrared Sensor

1. Introduction

During the last years, the World Health Organization (WHO) has defined that more than 3 million of people have died every year from a chronic respiratory disease (CRD), representing approximately 6% of global annual deceases [1]. A CRD is a disease of the airways, where the most commons are the asthma, chronic obstructive pulmonary disease (COPD), among others. One of the principal risk factors is the air pollution in the cities, occupational chemicals, dust and the frequent respiratory infections during childhood [2]. In recent years, several studies have analysed how asthma; a CRD; is exacerbated by pollutants [3], such as ozone (O_3), particulate matter (PM) with aerodynamic diameters less than 10 mm or 2.5 mm (PM10 or PM2.5, respectively), nitrogen dioxide (NO_2), carbon monoxide (CO) and sulphur dioxide (SO_2). Concerning this, the study of environmental parameters is very important considering the direct and indirect relationship between the climate, the environment and the respiratory health [4]. One of the most affective alternatives to obtain environmental and climate variables is the use of satellite remote sensing (RS) data. RS data have the major advantage of providing synoptic and frequent overviews of the Earth's surface, whereas the distribution of ground-based measurements is usually sparse and uneven. Additionally, using these data avoids expensive and time-consuming monitoring campaigns. These data can provide information related to vegetation, land use, temperature, air pollutants and others [5,6].

National Aeronautics and Space Administration (NASA) within the Earth Observing System (EOS) program have coordinated a series of satellite missions for global observations including the land surface (e.g., surface temperature, soil moisture, vegetation cover, and land use) observation [7]. EOS includes some important satellites as Terra, Aqua, Landsat-7 and Landsat-8. Terra and Aqua satellites were launched in 1999 and 2002, respectively. Their instruments include the Advanced Spaceborne Thermal Emission and Reflection (ASTER) and the Moderate Resolution Imaging Spectroradiometer (MODIS) [8]. MODIS is an instrument that acquires data in 36 spectral bands with different spatial resolution (from 250 to 1000 meters). This low spatial resolution can be considered a limitation in the analysis of medium scale cities. However, this sensor is able to obtain information of the entire Earth's surface every 1 to 2 days [9]. Landsat-7 and Landsat-8 are the last satellites from the Landsat Program launched in 1999 and 2013, respectively. Landsat-7 includes an Enhanced Thematic Mapper Plus (ETM+) sensor, while Landsat-8 includes two sensors: the Operational Land Imager (OLI) divided into 9 bands with 30 meters of spatial resolution (15m for panchromatic

band) and the Thermal Infrared Sensor (TIRS) instrument divided into 2 bands with 100 meters in native spatial resolution and resampled to 30 meters. OLI sensor also includes a Cirrus Band (B9) and a quality band (QA). B9 provides data of thin cloud contamination, while QA band evaluates the quality of each image pixel [10].

In the case of use satellite RS data to environmental and health studies, the most common satellites are from EOS program, with the main advantages related to the free access and the easiness to download. Typically, the use of satellite RS data is related to the retrieving of vegetation parameters, land use/cover and climate variables. Some of these variables are related to the use of spectral indexes as normalized vegetation difference index (NVDI), enhanced vegetation index (EVI), soil-adjusted vegetation index (SAVI), land surface temperature (LST) and others [11–13]. On the other hand, the air pollution has a big influence into the probability to get a CRD. The most common air pollutants are measured/quantified in the cities by an automatic air quality network (AQMN). These networks are implemented in order to establish a monitoring system in the cities, considering that these air pollutants have a high influence in the incidence of some CRDs and other diseases [14–17]. One of the approaches to relate air pollutants with RS is the Aerosol Optical Thickness (AOT) [18–20]. The AOT is a parameter that can be obtained from MODIS Aerosol product or Aerosol Optical Deep (AOD) ground stations (called AERONET), which allows to obtain measures of aerosols related with the air pollutants. Thus, several studies use AOT in order to retrieve air pollutants using RS data [21,22].

Regarding this, several studies show an increment in the use of RS data in health studies, related to environmental parameters [23,24]. These studies involve infectious disease epidemics and others CRDs, as asthma [25]. *Ayres-Sampaio et al.* [26], developed a study to evaluate the relationship between asthma hospital discharge and several environmental variables, in Portugal mainland, using RS data and spatial modelling. A set of five environmental variables were considered: near-surface air temperature (Ta) from the temperature profile of the MODIS sensor; relative humidity (RH) from meteorological station data interpolated by kriging method; vegetation density from MODIS NDVI product; and space-time estimates of nitrogen dioxide (NO_2) and particulate matter less than 10 mm (PM10), both from Land Use Regression (LUR) models based on data from AQMN stations. Districts were aggregated into three groups based on their percent urban cover, and the municipality was chosen as the sampling unit to assess the relationship between asthma hospital admission rates and environmental variables by season for the years 2003-2008. The results suggest that asthmatic people living in highly urbanized and sparsely vegetated areas are at a greater

risk of suffering severe asthma attacks that lead to hospital admissions. However, the limitations of this study are related to the global calibration, low spatial resolution of the RS data, atmospheric column effects, LUR statics models to derived air pollution and the dependence between some variables.

Alcock et al. [27] uses negative binomial regression model in order to relate some geographical variables with reductions in asthma hospitalisations, where one of the variables is the area-level data on vegetation. The study results showed that green spaces and gardens were associated with reductions in asthma hospitalisation when pollutants were lower. *Andrusaityte et al.* [28] identify the associations between neighbourhood greenness and asthma in preschool children, where the results show that an increase in the NDVI values data was associated with a slightly increased in the risk of asthma in children.

Fuertes et al. [29] identifies a non-consistent relationship between traffic-related air pollution (TRAP) on childhood asthma and allergic diseases documented during early-life persist into later childhood. One of the input variables to TRAP were the NO_2 and PM2.5 LUR from based on Corine land Cover (CLC) [30]. *Cillufo et al.* [31] used the CLC and NDVI as LUR inputs. The study showed that exposures related to greenness (measured by NDVI), greyness (measured by CLC) and air pollution are associated with respiratory general symptoms in schoolchildren.

The work presented in this thesis proposes an improvement and an update of different methodologies already cited in the literature [26] [27], but applied to a different geographical area (Quito, Ecuador), where the environmental conditions are extremely different and low probability to have RS data cloud free (high density) during all the year [32] is a reality. Moreover, this work aims to establish the most adequate spatial model to retrieve the hospital discharge of CRDs between 2013 and 2017 with a fine spatial resolution (30 meters). Thus, the study purpose: (i) to recover the more quantity of RS data (high cloud density) for Quito city [33]; (ii) to evaluate the most adequate RS data for the study area in order to retrieve air pollution variables [6]; (iii) to evaluate different techniques to select the most representative RS data and environmental variables predictors according to air pollution and health data [34] and; (iv) to compare several machine learning techniques (MLT) in order to model the CRDs in the urban area of Quito. The model chosen will be used for spatial mapping of the CRDs. Thus, this model will allow to identify the areas with more CRDs, getting some conclusions about the applicability of the model in order to explain a possible trend. The main idea is to find new alternatives in the use of RS data to have additional and useful answers about respiratory health.

1.1 Aim and objetives

The main objective of this work is to evaluate the applicability of RS data (processed for local conditions in Quito, Ecuador) in the study of CRDs, by the computation of the most adequate spatial models to retrieve hospital discharge of CRDs between 2013 and 2017. To achieve this goal, the following main steps have been applied (in their corresponding order):

1. Evaluating and improving the application of different methodologies to remove the effects of high-density clouds in order to have more RS data available for the computation of environmental indexes.
2. Investigate the most adequate RS data to use in the scope of this work, and their respective calibration and validation in Quito conditions. Several satellite sensors were investigated, e.g., MODIS, Landsat-7 ETM+, Landsat-8 OLI.
3. Developing different LUR algorithms to retrieve the environment variables from RS data, selecting adequately the predictors in order to model the air pollutants considering the sensor selected (previous step).
4. Studying the association between different CRDs and the environmental parameters retrieved from RS data, establishing spatial CRDs models from different MLT (Multiple linear Regression - MLR, Multilayer Perceptron - MLP, Support Vector Regression – SVR and Random Forest Regression - RFR).
5. Analyzing the limitations of this approach, defining the boundary conditions of the proposed model.

1.2 Thesis Outline

The core of this thesis is composed of six main chapters, as follows:

- The chapter 2 presents an overview of the theoretical subject about RS data and their applications in environmental and health studies. A perspective of the use of the different MLT in order to compute spatial models is also given.
- The chapter 3 presents an evaluation and an improvement of the Automatic Cloud Removal Method (ACRM) algorithm [35] to remove thin clouds considering Landsat-8, in order to recovery RS data and after to compute spectral indexes as NDVI. Thus, an automatic removal cloud method based on the cirrus band from Landsat-8 is proposed for the study area. This work was published in the “Remote Sensing Applications: Society and Environment journal” by Elsevier [33].

- The chapter 4 presents the evaluation of three different RS datasets in order to retrieve PM10 variable, considering a LUR model using only RS data from Landsat-8. This was published in “Environments” by MDPI [36].
- The chapter 5 presents a published paper in the “Environmental Monitoring and Assessment” journal by Springer, which is focusing in the selection of predictors in a LUR model, testing different MLT in order to retrieve O_3 concentration [34].
- The chapter 6 shows a development of the final spatial model. This work was submitted for the “International Journal of Environmental Research and Public Health” by MDPI. Thus, the evaluation of different MLT to compute a LUR model of the hospital discharge of CRDs in Quito, Ecuador is realized to achieve the thesis goals.
- Finally, the chapter 7 includes the discussion, conclusions and future work, identifying the achievable goals, the opportunities and the limitations of this study.

1.3 Study area

The study area is the urban area of Quito, the capital of Ecuador (Figure 1.1). The city has some special characteristics related to geology, climatology and location. Quito is crossed by the equatorial line in the North side. The study area latitude ranges between 0º30’S to 0º10’N and its longitude ranges between 78º10’W to 78º40’W. These coordinates delimit most of the urban zone, which is divided into 45 urban parishes. In the urban area is placed the downtown, and consequently higher air pollution concentration and high population density.

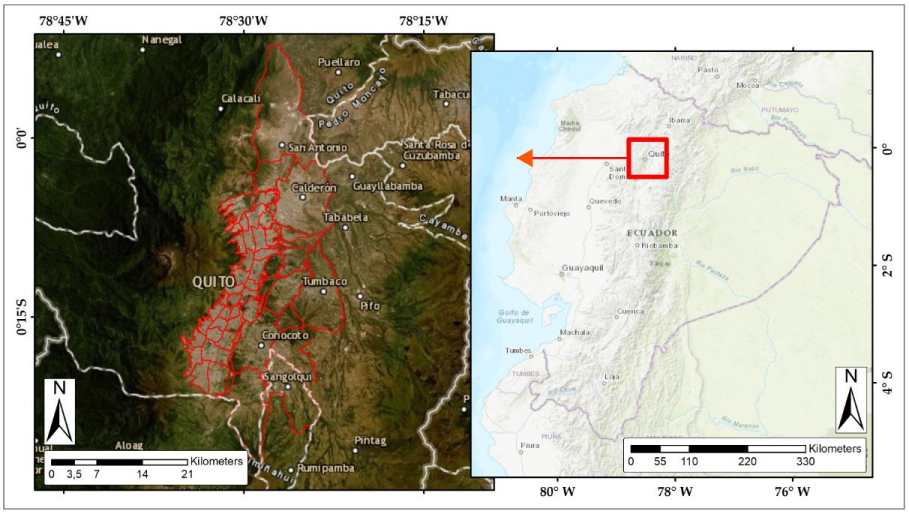


Figure 1.1. Study area. In the left image, the red polygons are the urban parishes in Quito.

According to the geology of northeastern of Ecuador, Quito is greatly influenced by the tectonic mechanisms responsible for the development of the Andes Mountains (Figure

1.2). Thus, some geological processes such as landslide, volcanism, erosion, weathering are presented in the city [37].

Figure 1.2. Quito in the middle of the Andean Region (adapted from The University of Texas. 25 July 2019, // utdirect.utexas.edu/apps/abroad/student/pgm_list/detail/nlogon/376/).

The high cloud density over Quito is very significant, all over the year. The specific reason is the influence of a high Andes Mountains region, situated in a tropical zone. The city elevation is approximately 2800 meters above sea level. Another of this area characteristic is the nonexistence of the traditional four seasons. The city has only one dry season and one wet season (February to May). The mean temperature during all the year is between 14 to 16 degrees Celsius (Figure 1.2).

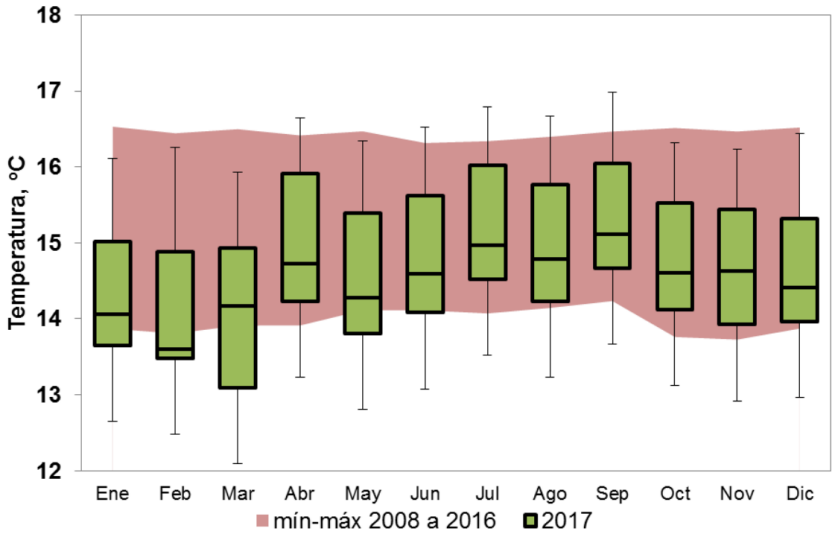


Figure 1.2. Air temperature average in Quito, Ecuador (adapted from [38]).

Another important climatology parameter is the solar irradiance. It is higher during August and September over 240 W/m^2 and the minimum solar irradiance is presented during the wet season with values lower than 160 W/m^2 (Figure 1.3).

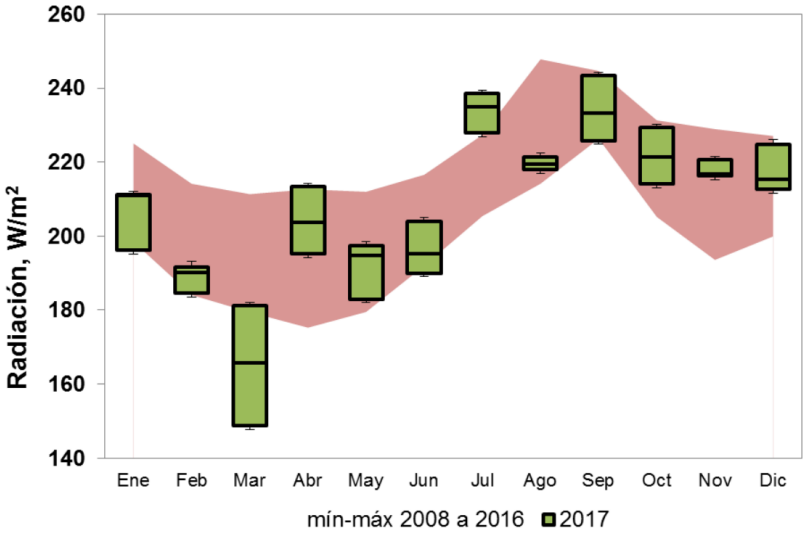


Figure 1.3. Solar irradiance average in Quito, Ecuador (adapted from [38]).

2. Theoretical Background

2.1 Sensor and platforms

RS can be defined as a technique to collect information about an object without making physical contact with it [39]. Satellite RS has the major advantage of providing synoptic and frequent overviews of the Earth's surface, whereas the distribution of ground-based measurements is usually too scarce and uneven to obtain enough information. The principles of the satellite RS could be defined in six stages: (i) an energy source, which produces the electromagnetic radiation to be captured by the sensor. The Sun is generally the energy source of passive sensors; (ii) the Earth's surface, which receives the incidence of the energy source; (iii) the platform and sensor ; (iv) the ground system, which receive the data; (v) the processing and analysis of the RS data and; (vi) the end users (Figure 2.1) [40].

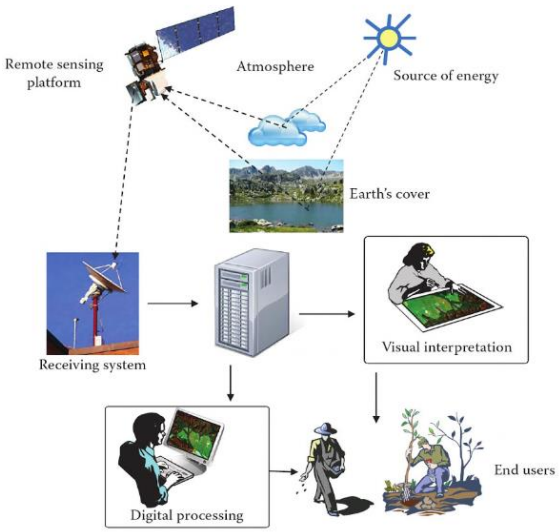


Figure 2.1. Main stages related to satellite RS (adapted from [40]).

The most significant advances in RS date back to the late of the 1960s, when NASA began the EOS program. EOS launched the first Earth Resources Technology Satellite (ERTS-1) in 1972 (renamed as Landsat-1). This date was a break point in the advance of the satellite RS, being the beginning of more than forty continuous years of Earth observation (EO) with the Landsat program. Landsat program is a series of EO satellite missions developed and supported by NASA [41]. The current orbit mission platforms to collect data are the Landsat-7 with the ETM+ sensor and Landsat-8 with the OLI and TIRS sensors (Figure 2.2). The next generation (Landsat-9) is expected to be launched in 2020.

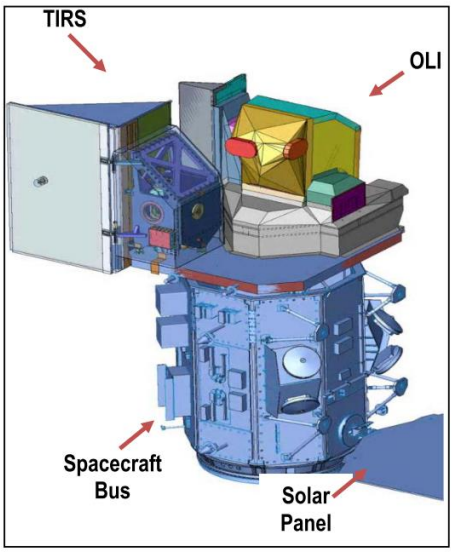


Figure 2.2. Illustration of Landsat-8 satellite (adapted from [10]).

EOS program has others EO satellites, as Terra and Aqua, launched in 1999 and 2002, respectively. These satellites are identical and both include the MODIS sensors on board [8]. The difference between both satellites are that Terra passes across the equator in the morning (North to South), while Aqua passes in the afternoon (South to North). One of the main advantages of these EOS satellites is the availability of the data (public) and they are free to download. The main characteristics of these EOS sensors and platforms are described in Table 2.1.

Table 2.1. Main characteristics of satellites and sensors considered in this work.

Satellite platform	Sensor	Bands (B)	Temporal resolution	Spatial resolution
Landsat-7	ETM+	B1 - Blue B2 - Green B3 - Red B4 - Near Infrared (NIR) B5 – SWIR 1 B6 - Thermal Infrared (TIR) Low Gain / High Gain B7 – SWIR 2 B8 – Panchromatic	16-days	30 m (B1-B5, B7) 100 m (B6) 15 m (B8)
Landsat-8	OLI TIRS	B1 - Coastal aerosol B2 - Blue B3 - Green B4 - Red B5 - Near Infrared (NIR) B6 - SWIR 1 B7 - SWIR 2 B8 - Panchromatic B9 – Cirrus B10 - Thermal Infrared (TIR) 1 B11 - Thermal Infrared (TIR) 2	16-days	30 m (B1-B7, B9) 15 m (B8) 100 m (B10-B11)

Terra Aqua	MODIS	36 spectral bands with different spatial resolution	1 to 2 days	250 m (B1–B2) 500 m (B3–B7) 1000 m (B8–B36)

According to the different characteristics in the sensors and platforms, it is important to evaluate the advantages and disadvantages of each RS data. For example, MODIS sensor is adequate for regional or global studies, where the spatial resolution is not a limitation. However, it is not so adequate in local scale studies, where the pixel size directly affects the results. An alternative is the use of Landsat-7 or Landsat-8 products. Nevertheless, it is important to emphasize that Landsat-7 sensor has a problem since 2003 in the Scan Line Corrector (SCL-Off) [42]. Moreover, one of the main advantages of Landsat satellites is the continuous data since 1972 to present (Figure 2.3).

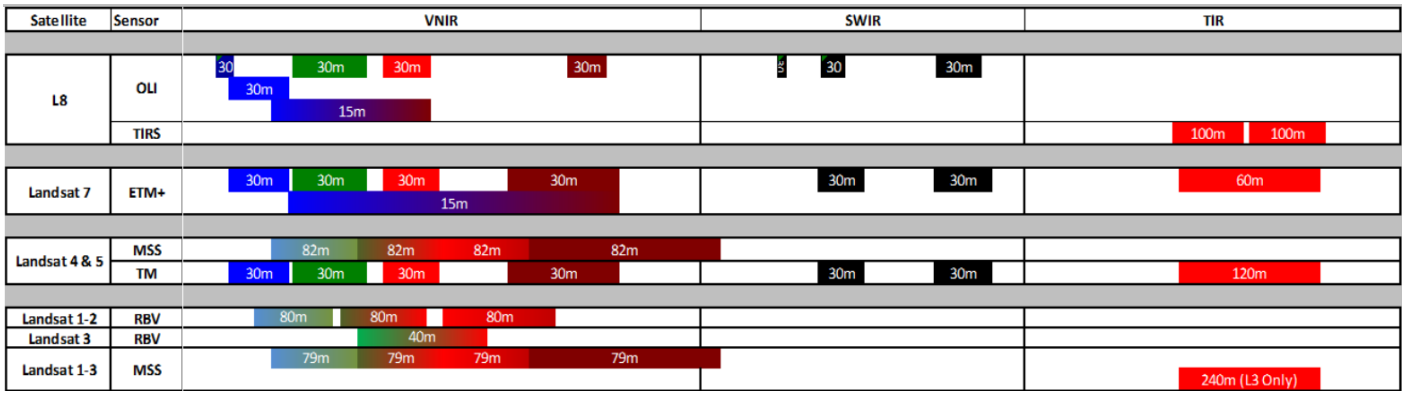


Figure 2.3. Landsat Missions multispectral data (adapted from [10]).

2.2 Data pre-processing

In order to have ready to use RS data products, several pre-processing steps must be applied. Thus, the geometric, topographic, radiometric and atmospheric corrections are mandatory in the use of RS data, because RS raw data give the surface radiance in the form of Digital Number (DN). The DNs must be converted to physical units, correcting all the possible effects. The explanations and sequence of the corrections (Figure 2.4) are explained below:

- The geometric and topographic corrections are necessary in order to repair the geometric deformations. This distortion needs to be corrected finding the geographical reality on the ground, associating to a coordinate reference system, ground control points, altitude, etc. [43].
- The radiometric correction reduces the influence of inconsistencies in image brightness values, which could limit the analysis of RS data [44]. In this correction, DN is converted into radiance. Then, the radiance is converted into

top of the atmosphere (TOA) reflectance data and in brightness temperature (BT) in the thermal bands.

- The atmospheric correction allows to remove the atmospheric effects due to absorption and scattering effects. Several algorithms can be used to estimate the surface reflectance. One of the most popular and simplest method is the empirical Dark-Object Subtraction (DOS) [45]. DOS assumes that the reflectance of dark objects has a considerable component of atmospheric scattering, searching the darkest pixel value to each band. More complete physical methods are Second Simulation of a Satellite Signal in the Solar Spectrum Vector (6SV) and Fast Line-of-sight Atmospheric Analysis of Spectral Hypercubes (FLAASH). 6SV can simulate the solar radiation on the ground and in the atmosphere under a variety of conditions of both ground surface and atmosphere [46]. However, it requires several local data, as meteorological variables. FLAASH is an ENVI software package based on MODerate resolution atmospheric TRANsmission (MODTRAN) radiation transfer code [47]. FLAASH uses physics-based derivation of atmospheric properties such as surface pressure, water vapor column, aerosol and cloud overburdens in order to convert TOA reflectance into surface reflectance values [47,48]. The selection of the atmospheric correction method will depend on the availability of the data for the study area. Some studies show that the physical methods are more accurate than empirical methods [49].

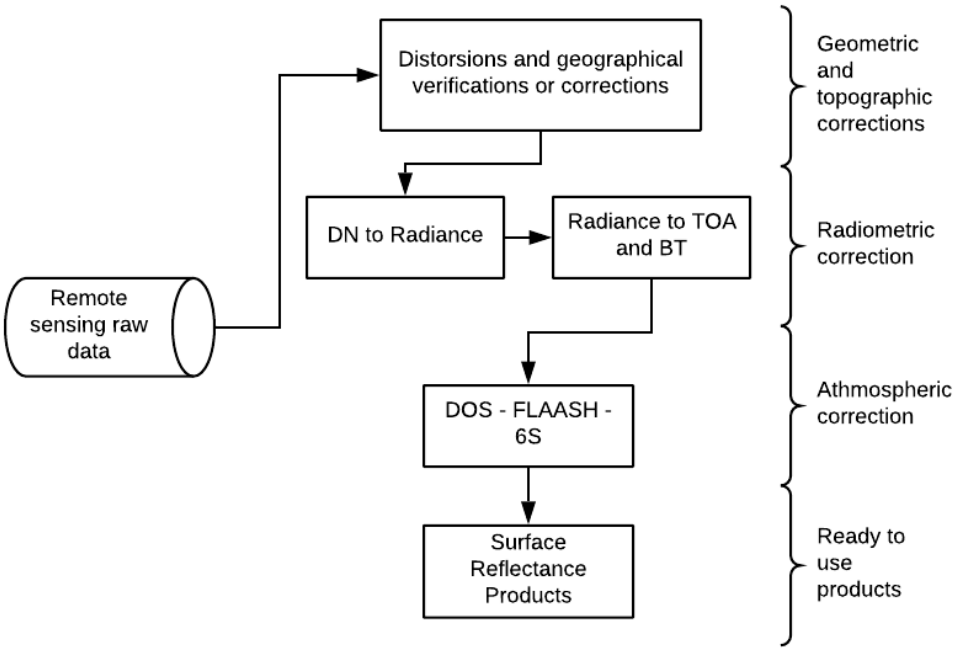


Figure 2.4. Workflow of RS data pre-processing tasks.

Landsat-7 and Landsat-8 can download in Level 1 LT1 product, available on United States Geological Survey (USGS) website (http://earthexplorer.usgs.gov). This product is already radiometrically calibrated and orthorectified, avoiding the geometric and topographic calibration. Thus, only radiometric and atmospheric correction must be applied. On the other hand, the USGS provides Landsat surface reflectance Level-2 products (L2T). L2T products are radiometric and atmospherically corrected, where the products include surface reflectance and BT ready to use. In order to obtain L2T products, Landsat-7 uses the Landsat Ecosystem Disturbance Adaptive Processing System (LEDAPS) [50], while, Landsat-8 uses the Landsat Surface Reflectance Code (LaSRC) [51]. Moreover, L2T products are available to download from the Earth Resources Observation and Science (EROS) Center Science Processing Architecture (ESPA) at the demand interface (https://espa.cr.usgs.gov/). In the case of MODIS sensor, the MOD09 (Terra) and MYD09 (Aqua) are the products used in this work. They derived the surface reflectance [52]. These products provide images at ground level, without atmospheric scattering or absorption.

2.3 Data processing

RS information about vegetation, temperature or land use are extremely related to useful applications in the areas of environmental monitoring, climatology, biodiversity conservation, agriculture, forestry, urban green infrastructures, air pollution and other related fields [53]. In most of these applications, RS is used to acquire surface information through spectral indexes. These indexes are the result of processing the surface reflectance data and BT.

The NVDI is one of most popular vegetation spectral indexes. It provides information about health vegetation [54], where high NDVI values correspond to dense or primary vegetation (usually higher than 0.3), and low values can correspond to sick vegetation or indicate the presence of bare soils. Negative values correspond to water or snow. NDVI is computed using the surface reflectance data from the NIR and RED bands (Equation 2.1):

$$NDVI = \frac{NIR - RED}{NIR + RED} \tag{2.1}$$

Moreover, SAVI is an improvement of NDVI. It considers a soil correction factor – LS (usually LS = 0.5) [55]. LS minimizes the soil brightness influences, especially, when urban areas with low vegetation cover and bare soils exist in the scene (Equation 2.2):

$$SAVI = (1 + LS)\frac{NIR - RED}{NIR + RED + LS} \tag{2.2}$$

The EVI enhances the vegetation (Equation 2.3) in areas with high biomass, as forests. It improves vegetation monitoring through a de-coupling of the canopy background signal and a reduction in atmospheric influences [56].

$$EVI = G * \frac{NIR - RED}{NIR + C1 * RED - C2 * BLUE + L} \quad (2.3)$$

where G is the gain factor (2.5), L is the canopy background adjustment (1), C1 (6) and C2 (7.5) are coefficients for atmospheric resistance. The Red and NIR bands in this index allowed to detect built-up areas and bare lands areas [57].

One important parameter related to surface energy and water balance is the LST [58]. The LST is the relative temperature of the land surface computed from RS data. It is computed from the TOA BT (TIRS bands), in Kelvin. The Equation 2.4 allows to compute the LST in degrees Celsius.

$$LST = \frac{BT}{\left(1 + \left(\frac{\lambda * BT}{p}\right) ln\mathrm{E}\right)} - 273.15 \quad (2.4)$$

where λ is the centre wavelength (10.8 μm), E is the emissivity obtained from the Equation 2.6; p is estimated using Equation 2.5, where h is the Planck constant (6.626e-34 Js), c is the speed of light (2.998e8 m/s), and s is the Boltzmann constant (1.38e-23 J/K).

$$p = \frac{h * c}{s} \quad (2.5)$$

The Equation 2.6 was used to compute the emissivity (E) [59], where E is the efficiency that a surface emits heat as TIR radiation [60].

$$E = \begin{cases} E_s, & NDVI < NDVI_s \\ E_s + (E_v - E_s)P_V, & NDVI_s \leq NDVI \leq NDVI_v \\ E_v, & NDVI > NDVI_v \end{cases} \quad (2.6)$$

where E_s and E_v represent the E in the soil and vegetation, respectively. $NDVI_v$ and $NDVI_s$ are the NDVI in vegetation and soil, respectively. P_V is the proportion of vegetation in the study area computed based in the Equation 2.7.

$$P_V = \left(\frac{NDVI - NDVI_s}{NDVI_v - NDVI_s}\right)^2 \quad (2.7)$$

The Landsat L2T products have available to download all the indexes presented in the previous equations, except the LST. MODIS provides MOD13 and MYD13 products in (NDVI and EVI data ready products, respectively). Moreover, MOD11 and MYD11 products provides LST ready to use.

2.4 Remote sensing in environmental and health studies

For environmental scientists, the most important aspect of RS data is to provide relevant information for monitoring Earth's resources. The benefits of environmental monitoring of RS data in comparison with other methods, are the global view over of the Earth's surface, the multiscale observations (regional to local studies), the possibility to repeat observations (very useful in temporal studies), the immediate transmission (real-time transmissions in some cases) and the facility to combine with other geographical information [40]. Thus, several studies contain RS, environmental and health data involving monitoring, spatial predictive modelling, surveillance, and risk assessment [23]. These studies are specifically associated with the retrieving of air pollutants (PM10, PM2.5, O_3, NO_2, SO_2) in combination with ground data and other geographical variables as vegetation, knowing the possible correspondence with some vector-borne [61] and respiratory diseases [62–65]. Several works investigate this relationship. For instance, *Liang et al.* [62] established AOD-PM2.5 models to study the spatial correlation with allergic rhinitis in Taiwan. The study found a high correlation between these two factors (AOD-PM2.5 and allergic rhinitis), particularly in spring and fall. *Al-Hamdan et al.* [63] shows in their study the important relationship between PM retrieved from MODIS AOT and the respiratory system cancer. *Ai et al.* [64] presents in this study, the relationship between air pollution and asthma cases, where RS data were used to estimate the yearly mean of air pollutants. Additional studies, as *Andrusaityte et al.* [28], use vegetation multispectral indexes. They identified the associations between neighbourhood greenness and asthma in preschool children, where the results showed that the increase in the NDVI was associated with a slightly increased of the relative risk of asthma in children. In opposition to these works, *Li et al.* [65] founded that NDVI does not have an association with respiratory and allergic outcomes. They concluded that living closer to green parks appeared to be a risk factor for asthma.

2.5 Ground and health data

2.5.1 Air pollutants and meteorological measurements ground data

In order to compute spatial models to retrieve environmental and health variables, ground data are necessary to calibrate the models, specifically air pollutants and meteorological variables. Thus, a valid alternative to collect daily ground measurement data is through an air quality monitoring network (AQMN). An AQMN is a network with a series of fixed stations equipped with sensors to measure some air pollutants, such as PM10, PM2.5, O_3, NO_2, CO and SO_2 (Figure 2.5). The AQMN stations give an

understating about air pollution and the impact over the human health. Some AQMN have meteorological sensors (MD) that allow to obtain field measurements of pressure, wind direction, relative humidity, precipitation, wind speed, air temperature and solar irradiance.

A good planning of the location of the AQMN stations is mandatory. However, in most of the cases a high maintenance cost by station [66], a low quantity of stations in large cities or non-representative spatial distribution [67] are the main problems.

The AQMN available in the study area (Quito, Ecuador) is the “Red Metropolitana de Monitoreo Atmosférico de Quito” (REMMAQ)[38]. It has worked since 2002 with nine monitoring stations with air pollutant and meteorological sensors (Table 2.2).

Figure 2.5. AQMN station in Quito, Ecuador (adapted from [38]).

Most of the data retrieved by REEMAQ have influence in the respiratory health. The data is available to download in the Environmental Secretary of Quito web page (http://www.quitoambiente.gob.ec/ambiente/index.php/datos-horarios-historicos) for free.

Table 2.2. Field sensors of the REEMAQ by station

Station	Variables measured
Cotocollao	PM2.5, SO_2, CO, O_3, NO_2, PM10, MD
Carcelen	PM2.5, SO_2, CO, O_3, NO_2, PM10, MD
Belisario	PM2.5, SO_2, CO, O_3, NO_2, MD
Jipijapa	PM2.5, SO_2, CO, O_3, NO_2, PM10, MS
Camal	PM2.5, SO_2, CO, O_3, NO_2, MD
Centro	PM2.5, SO_2, CO, O_3, NO_2
Guamani	SO_2, CO, O_3, NO_2, PM10, MD
Tumbaco	SO_2, O_3, PM10, MD
Los Chillos	PM2.5, SO_2, CO, O_3, NO_2, MD

It is important to measure the air pollutants in the cities. In Quito, the PM2.5 is one of the air pollutants over the WHO limits during each year (Figure 2.6). Like it, we have

more air pollutants over the WHO limits, which should be monitored in order to obtain a better air quality.

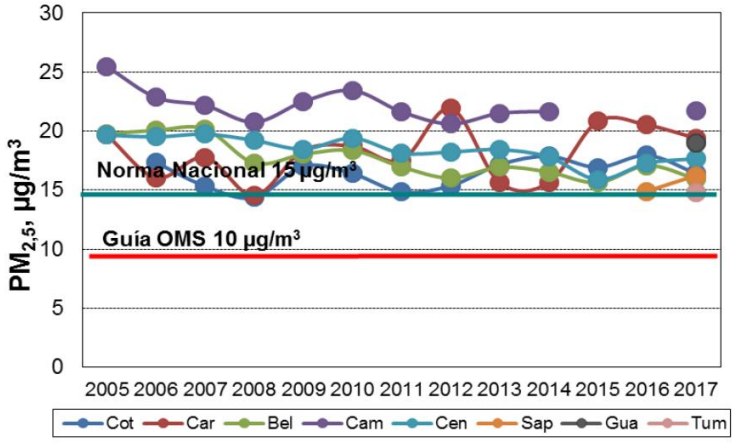


Figure 2.6. PM2.5 average measures during the years 2005 to 2017 in each AQMN station (adapted from [38]).

2.5.2 Respiratory health data

The respiratory health data are significant considering the hospital discharge by a CRD. According to the WHO, a CRD is a disease of the airways and other structures of the lung. The most common are asthma, chronic obstructive pulmonary disease (COPD), occupational lung diseases and pulmonary hypertension [1]. A hospital discharge is defined as the patient who has stayed at least one night in the hospital, including dead people during the health care. As already referred, one of the principal risk factors to get a CRD is the air pollution in the cities [2], considering the exacerbating by air pollutants [3] and meteorological conditions.

In this work, the CRDs were filtered in the ICD-10 codes: J40 – J47. This classification is based in the International Classification of Diseases 10 (ICD-10) from the WHO [68]. The codes J40-J47 includes diseases as asthma and bronchitis.

In the case of the study area, the National Institute of Statistics and Census (INEC) is the official government institution in charge to provide the information about population and other socioeconomic statistic variables in Ecuador. This information is public in a parish ("*parroquia*") scale. One of the variables provides by INEC is the hospital discharge information. It is available to download from INEC web page (http://www.ecuadorencifras.gob.ec/camas-y-egresos-hospitalarios/).

2.6 Models

The base fundamentals of an empirical LUR model is considered in this research. In this work, a LUR model is a regression which uses the air pollutant ground measurements as dependent variable and other geographic variables as independent variables (traffic,

roads, land use, topography, etc.) in a multivariate regression model [69]. In most of the cases, the MLR is used to compute the LUR model [70]. However, one limitation is the use of some static geographic variables, as the distance to roads, traffic count, land use, etc., mainly when the geographical variables are not updated. The classical LUR model computes spatial air pollutants and then are compared with health data. This study aims to establish a spatial model-based on an empirical LUR model, considering the CRDs as the dependent variable and other dynamic geographic variables (RS, air pollution, meteorological parameters) as independent variables. In order to compute the LUR models, some MLT can be applied and after compared in order to find the most effective algorithm.

2.7 Machine learning techniques

Machine learning is a category of algorithms that receive input data and use statistical analysis to predict an output while updating outputs as new data becomes available [71]. MLT uses statistical and computational methods to learn information from the dataset without being based on a predetermined equation as a model. The algorithms adaptively improve their performance as the number of samples available for learning increases [72]. MLT algorithms range from the simplest linear regression models until the more complex algorithms, as a neural network (Figure 2.7).

MLT has two kinds of learning: unsupervised and supervised learning. The aim of the first is the regularization of the input data through clusters and not in the output data (without supervision). The second one has a supervision process in both data (input and output). In this work, MLTs with supervised learning were adopted. Thus, classification and regression processes are a supervised learning problem where there is an input x (independent variables) and an output y (dependent variable). The objective of the MLT is to predict the output, considering a learning process (Equation 2.8).

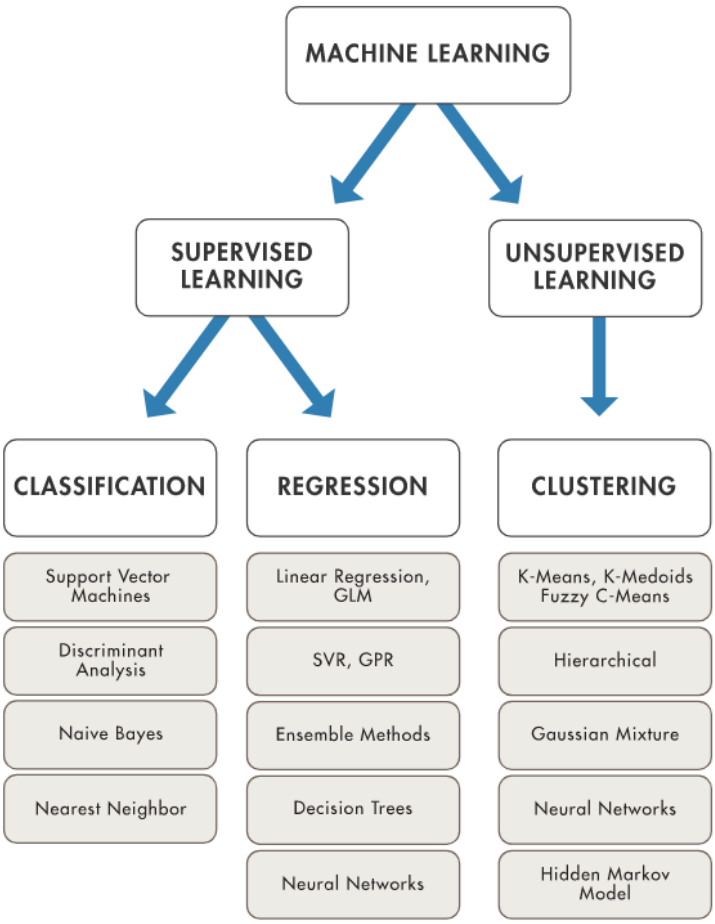


Figure 2.7. MLT (adapted from [72]).

$$y = g(x|\theta) \tag{2.8}$$

Where $g(.)$ is the regression function (in regression) or the discriminant function (in classification), θ are the parameters or independent variables. Y is the dependent variable (a number in regression and a class code in classification).

2.7.1. Multiple linear regression (MLR)

MLR also known as multiple regression, is a multivariate linear regression, and is considered the simplest MLT. It uses several explanatory variables (independent variables) to predict the outcome of a response variable (dependent variable), generating a model with a linear relationship. The linear regression is a weighted sum of several input variables [71] (Equation 2.9).

$$y = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon \tag{2.9}$$

Where y is the dependent variable, x_i are the independent variables, β_0 is the intercept, β_p are the slope coefficient for each explanatory variable and ε are the residuals.

2.7.2. Stepwise regression

Stepwise regression is a MLR, which has an automatic process of the selection of predictors (independent variables). It is a combination of the forward and backward techniques [73]. The forward selection begins with no candidate predictors in the model. Then, the variables are selected in function to the highest coefficient of determination (R^2), adding one by one variable. The backward selection is the opposite. It begins with a model considering all the predictors included and then they are excluded to test the highest R^2. The problem with the backward selection is that it may include variables that are not necessary (could present a high correlation) [74].

2.7.3. Partial Least Square (PLS)

PLS regression is a similar technique to principal components regression, which uses latent variables or components as predictors [75]. PLS projects the predictors and dependent variable into a new space in different hyperplanes or latent variables. The advantage to project to new latent variables is to avoid multicollinearity. The PLS regression is showed in Equation 2.9.

$$y = a_1 t_1 + a_2 t_2 + a_3 t_3 + \cdots a_n t_n \quad (2.9)$$

Where t_i are the latent variables or components. They are themselves linear combinations of the independent variables (xi), as presented in Equations 2.10, 2.11 and 2.12.

$$t_1 = b_{11} x_1 + b_{12} x_2 + \cdots b_{1p} x_p \quad (2.10)$$
$$t_2 = b_{21} x_1 + b_{22} x_2 + \cdots b_{2p} x_p \quad (2.11)$$
$$t_i = b_{i1} x_1 + b_{i2} x_2 + \cdots b_{ip} x_p \quad (2.12)$$

Additionally, PLS generate an orthogonal transformation to obtain components by finding the most appropriate model to explain the variance, starting from the maximise covariance matrixes [76].

2.7.4. Multilayer perceptron (MLP)

MLP is part of an artificial neural network (ANN). It is an MLT used to solve problems in classification and regression. Moreover, MLP is based on the perceptron algorithm, which takes an input dataset, then aggregates it with a weighted sum and finally, it returns 1 only if the aggregated sum is more than some specific threshold or if not returns 0 (Figure 2.8). The Equation 2.13 shows the decision rule of the multilayer perceptron algorithm.

$$y = 1 \text{ if } \sum_{i=0}^{n} w_i * x_i \geq 0 \quad (2.13)$$

$$y = 0 \text{ if } \sum_{i=0}^{n} w_i * x_i < 0$$

Where, x_i are the predictors and w_i are the weights of each variable.

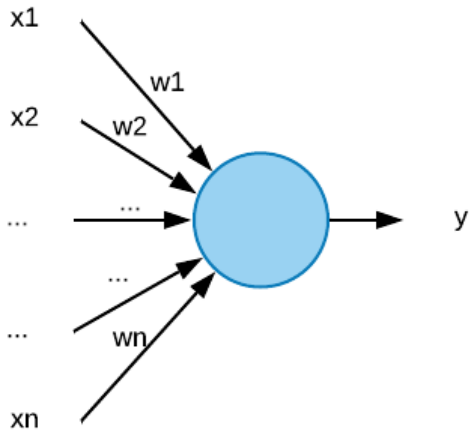


Figure 2.8. Perceptron algorithm schema.

The MLP uses a series of neuronal activities where the ideal is to have an interconnection weights in a non-linear multilayer perceptron [77,78]. The simplest MLP has three-layers (Figure 2.9). The first layer is the input layer and the last is the output layer; the middle layer is the hidden layer. This architecture is used in regression problems. However, the number of hidden layers in an MLP and the number of nodes in each layer can have variations according to each problem. Thus, more nodes give more sensitivity, but a high risk of overfitting [79].

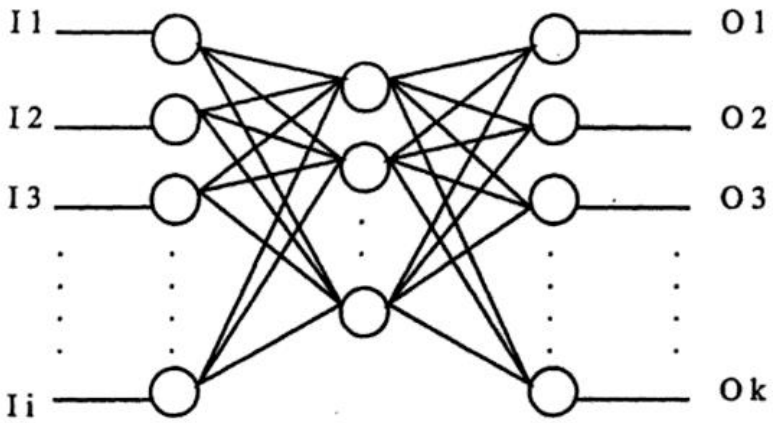


Figure 2.9. The simplest MLP architecture (one input layer, one hidden layer, one output layer). Figure adapted from [78].

The MLP uses a backpropagation method in order to train the model. The backpropagation methodology identifies if the MLP has an error in the prediction.

2.7.5. Support Vector Regression (SVR)

SVM was developed to solve classification problems, extending to regression problems (SVR). SVM transforms nonlinear regression into a linear regression with the transformation between the original low dimensional input space into a high dimensional feature space using kernel functions. These kernel functions carry a low dimensional plane to a higher dimensional space to separate the variables using a hyperplane. Thus, decision vectors were obtained (Figure 2.10) [80]. In the new higher dimensional space, several linear models are constructed to obtain an optimal solution [81]. The SVM and SVR work into a higher-dimensional space. The main difference is that the SRV uses a continuous numerical variable as dependent variable [82].

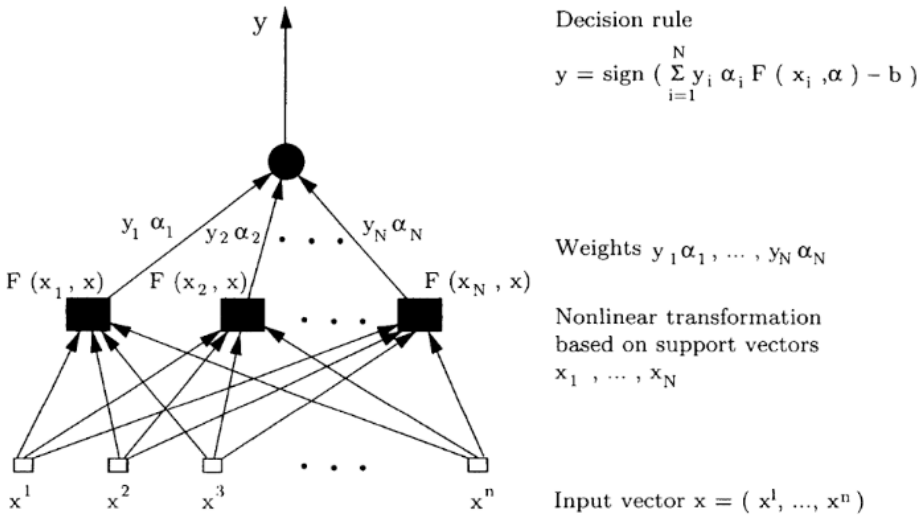


Figure 2.10. The two-layer SVM is a compact realization of an optimal hyperplane in the high-dimensional feature space Z. We pass from a complex non-linear function to a simpler linear function (adapted from [80]).

2.7.6 Random Forest (RF)

RF is an effective ensemble learning algorithm in classification or regression. It uses the training dataset to generate multiple decision trees (forest) being less sensitive to the overfitting problem through the bootstrap aggregation commonly called bagging. The bagging trains each decision tree on a different data sample, where the sampling is done with replacement [83,84]. The decision trees make a simply combining according to their weights in order to determine the final output (Figure 2.11). Moreover, RF is considered one of the most effective non-parametric ensemble learning methods in image analysis [85]. The Equation 2.14 shows the RF regression in a general form [86].

$$\hat{f}_{rf}^{K}(\mathrm{x}) = \frac{1}{k}\sum_{k=1}^{k} \mathrm{T}(\mathrm{x}) \tag{2.14}$$

Where, x is an input vector from the values of the different features analysed for a given training area. RF builds a number K of regression trees $\{T(x)\}_1^K$ averaging the results. In order to avoid the correlation between different trees, RF increases the trees with the different data subset created (bagging).

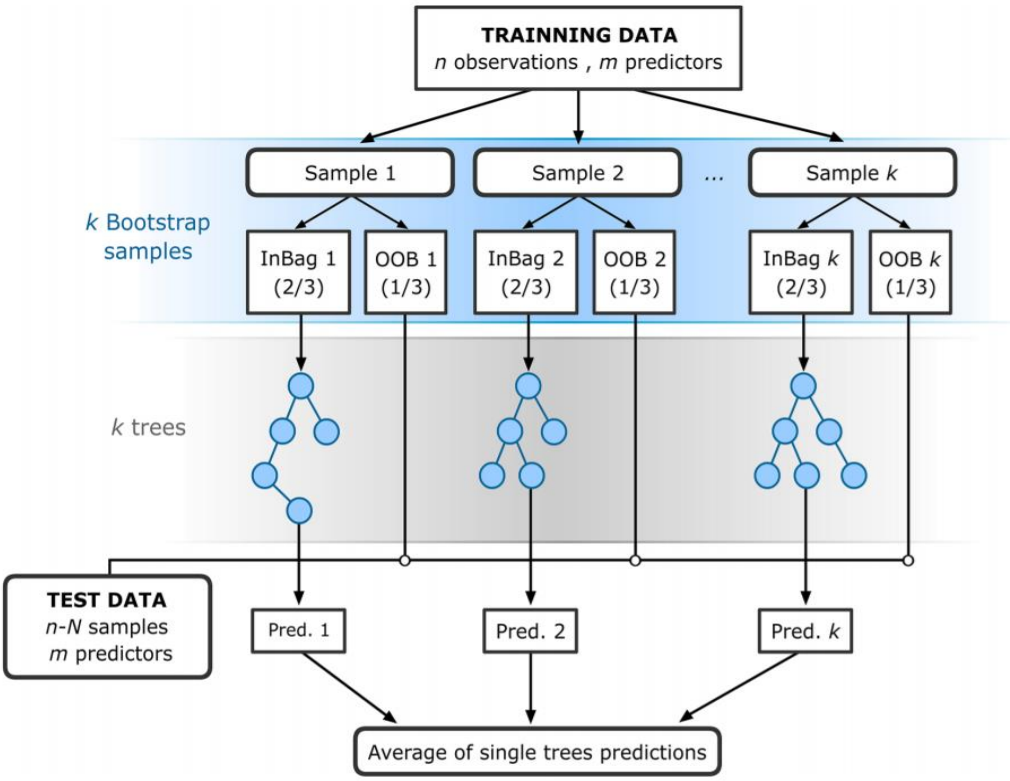


Figure 2.11. Flowchart of RF for regression. The RF receive the input training data, then RF builds a number k of regression trees with different training data subsets (Bagging). Adapted from [87].

3. Article 1: Improving NDVI by removing cirrus clouds with optical remote sensing data from Landsat-8 – A case study in Quito, Ecuador

Cesar I. Alvarez-Mendoza[1,2*], Ana Teodoro[1,3], and Lenin Ramirez-Cando[2]

[1] University of Porto, Department of Geosciencies, Environment and Land Planning, Faculty of Sciences, Rua Campo Alegre 687, Porto 4169-007, Portugal; calvarezm@ups.edu.ec

[2] Universidad Politécnica Salesiana, Grupo de Investigación Ambiental en el Desarrollo Sustentable, Carrera de Ingeniería Ambiental, Quito, Ecuador; lramirez@ups.edu.ec

[3] Earth Sciences Institute (ICT), Pole of the FCUP, University of Porto, Porto, Portugal; amteodor@fc.up.pt

Academic Editors: X.-L. Ding, G. Xian

Received: 7 June 2018 / Accepted: 13 November 2018 / Published: 16 November 2018

Journal: **Remote Sensing Applications: Society and Environment, Elsevier**

Volume 13, January 2019, Pages 257-274; doi.org/10.1016/j.rsase.2018.11.008

3.1 Abstract

The Andean region has a high cloud density throughout the year. The use of optical remote sensing data in the computation of environmental indices of this region has been hampered by the presence of clouds. To maximize accuracy in the computation of several environmental indices including the normalized difference vegetation index (NDVI), we compared the performance of two algorithms in removing clouds in Landsat-8 Operational Land Imager (OLI) data of a high-elevation area. The study area was Quito, Ecuador, which is a city located close to the equator and in a high-elevation area crossed by the Andes Mountains. The first algorithm was the automatic cloud removal method (ACRM), which employs a linear regression between the different spectral bands and the cirrus band. The second algorithm was independent component analysis (ICA), which considers the noise (clouds) as part of independent components applied over the study area. These methods were evaluated based on several images from different years with different cloud conditions. The results indicate that neither algorithm is effective over this region for the removal of clouds or for NDVI computation. However, after improving ACRM, the NDVI computed using ACRM showed a better correlation than ICA with the MODIS NDVI product.

Keywords: cloud removal, optical remote sensing, Landsat-8 OLI, Quito, NDVI

3.2 Introduction

Optical remote sensing (ORS) data have the major advantage of providing synoptic and frequent overviews of the Earth's surface, but the distribution of ground-based measurements is scarce in some parts of the world. ORS data include visible (VIS), short-infrared (SWIR), and thermal infrared (TIR) regions of the electromagnetic spectrum [88].

Regions with a high cloud density during most of the year, such as the Brazilian Amazon [39,89,90] and the Andean region [91], are particularly challenging for ORS, especially in terms of the computation of the environmental indices, such as normalized difference vegetation index (NDVI) [92,93]. Several studies on cloud density have been conducted based on Landsat data [39,89,90]. [94] takes the spectral/spatial characteristics of Sentinel-2 as a template for instruments with similar properties as Sentinel-2 to investigate the relevant cirrus effects. [95] proposed a method based on the classic homomorphic filter executed in the frequency domain to thin cloud removal for visible remote sensing images. [96] propose an empirical technique for the removal of thin cirrus scattering effects in OLI visible near infrared and shortwave IR spectral regions. In the

work of [97], the top-of-atmosphere reflectance of thin clouds is modeled using the empirical relationships of the deep blue and blue bands of Landsat-8 OLI.

The Landsat program has provided calibrated and high-resolution spatial data of the Earth's surface for more than 45 years. Landsat-8, launched in February 2013, is the latest satellite in a continuous series of land remote sensing satellites that began in 1972. Landsat-8 has provided data to support several fields and research topics, such as agriculture, forestry, geology, land use, air contamination [98], and the removal of clouds in remote sensing images [35,99–104]. Landsat-8 includes two sensors: the Operational Land Imager (OLI), which is divided into nine bands with a spatial resolution of 30 m, and the Thermal Infrared Sensor (TIRS) instrument, which is divided into two bands with a native spatial resolution of 100 m. The OLI bands include a cirrus band (B9). Cirrus clouds are high-altitude clouds in the atmosphere and are mainly composed of miniscule ice crystals [105]. They are strong reflectors of radiation at a wavelength of 1.38 μm [10]. Cirrus clouds have a significant number of thin, non-spherical ice crystals that can absorb sunlight and attenuate the pixel values of surface reflectance in remote sensing [106]. Additionally, cirrus clouds limit the accuracy in the computation of environmental indices. Thus, it is crucial to remove them [93].

The purpose of this work is to develop an approach to remove clouds and noise in optical remote sensing data without losing surface pixel accuracy in order to compute environmental indices, such as NDVI. Several methods have been tested to remove clouds considering Landsat-8 data in different places around the world with satisfactory results. Some of these methods used a reference Landsat-8 image to patch the cloudy area [99,100,107], or combine Landsat-8 with other sensors [108], or work with the Landsat-8 cirrus band (B9) [35,102,109]. All these studies were conducted in low elevation regions and in no tropical areas. Both parameters can have an effect over cirrus clouds, considering that these clouds can form at any altitude between 5.0 km and 14 km above sea level. In the tropical regions, cirrus clouds cover around 70% of the region's surface area.

In this work, to remove cirrus clouds over an area in the Andean region (Quito, Ecuador) considering the Landsat-8 cirrus band (B9), two methods were evaluated: the automatic cloud removal method (ACRM) and independent component analysis (ICA). ACRM was first tested on images of Sydney, Australia [35]. The algorithm applies a linear regression between each multispectral band and the cirrus band (B9), evaluates the coefficient of determination (R^2) and slope in some areas, and generalizes them for the entire image [35]. In order to remove clouds, the algorithm uses the area with the highest R^2 to extrapolate values for the entire image. In ICA, independent components (ICs) are

separated, and one of them is the component that storing the thin clouds [110]. This algorithm was tested on Landsat-8 images of a low elevation region (North Carolina, USA), and the results were satisfactory [102]. The performance of the two methods in removing clouds and their efficiency in future computation of environmental indices such as NDVI are evaluated based on the same image.

3.3 Materials and Methods

3.3.1. Study Area and Dataset

3.3.1.1. Study Area

The study area is Quito, the capital of Ecuador (Figure 3.1). The equator line crosses the city in the north part. The Quito latitude ranges between 0º30'S to 0º10'N and its longitude ranges between 78º10'W to 78º40'W. Quito has a high elevation of approximately 2800 m. The cloud density over the city is considerable, all over the year. Quito has only one dry season and one wet season, considering that it is a tropical zone and is influenced by the Andes Mountains. In 2015, the mean minimum and maximum temperatures were approximately 9.0°C and 25.4°C, respectively, with a high precipitation of approximately 1126 mm [111]. The geology of northeastern Ecuador and present-day physical processes related to geology are greatly influenced by the tectonic mechanisms responsible for the development of the Andes Mountains. Both geology and active physical processes (landsliding, volcanism, erosion, weathering) are complex and varied [37].

3.3.1.2 Dataset

In this study, ten Landsat-8 L1T images were processed to evaluate and improve the two methods to remove clouds. Seven images of Quito, Ecuador (Path 10; Row 60) from different years (Figure 3.2); one image of Pedernales, Ecuador (Path 11; Row 60), which is a coastal region with characteristics similar to those of Sydney; and the image of Sydney, Australia (Path 89; Row 83) used in [35] were considered.

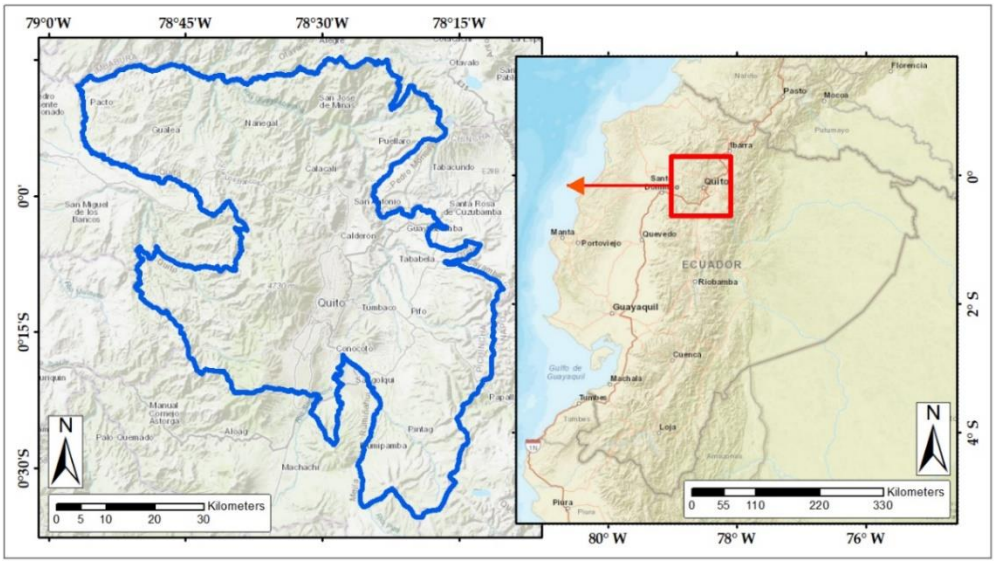


Figure 3.1. Quito Metropolitan Area

Images at the L1T processing level were considered because they take advantage of geometric and radiometric corrections [10]. Moreover, the MODIS MOD13Q1 product (tiles H10V08 and H10V09) for the study area was also used in order to compare the results obtained in the computation of NDVI (further details in Section 3.4) (Table 3.1).

Table 3.1. Characteristics of datasets used in this study

Sensor	Product	Spatial Resolution	Temporal resolution	Bands/Products
Landsat-8	L1T	30 m	16 days	Coastal aerosol, blue, green, red, near infrared, SWIR 1 and SWIR 2, Cirrus, Thermal Infrared 1, Thermal Infrared 2
MODIS	MOD13Q1	250 m	16 days	NDVI/EVI Values

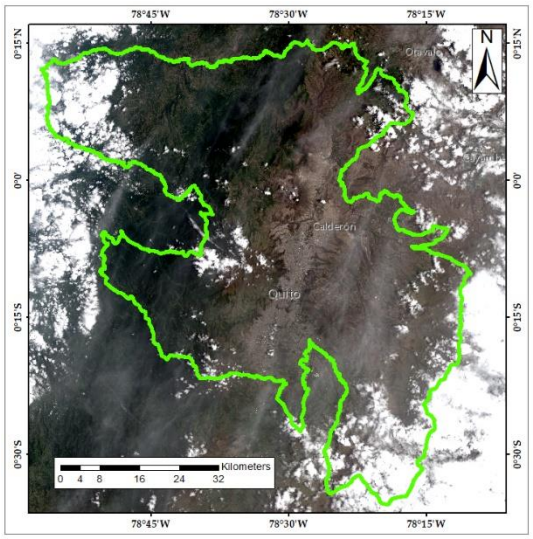
3.3.2.Methodology

Two methods to remove clouds, ACRM and ICA, were evaluated in this work for Landsat-8 images and the corresponding cirrus band (B9). Most of the processing steps were implemented in R programming language [112] and its associated packages: raster version 2.5-8 [113], rgdal version 1.1 [114], and gdalutilities version 2.0.1.7 [115]. Furthermore, ENVI® and ERDAS® software were used to perform some image processing tasks.

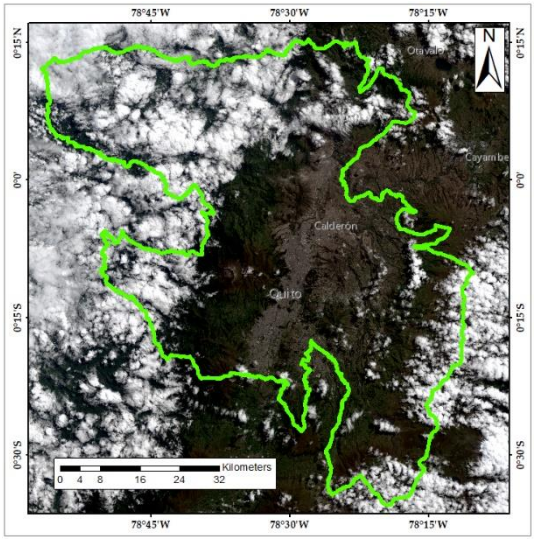
3.3.2.1. Automatic Cloud Removal Method (ACRM)

ACRM attempts to obtain clean pixel data from each digital number DN recorded at each OLI multispectral band $i = 1,2,3,4,5,6,7$. DN contains clean pixel data plus contaminated data at the location (u, v). Contaminated data are affected by clouds [35]. The model can be expressed as follows:

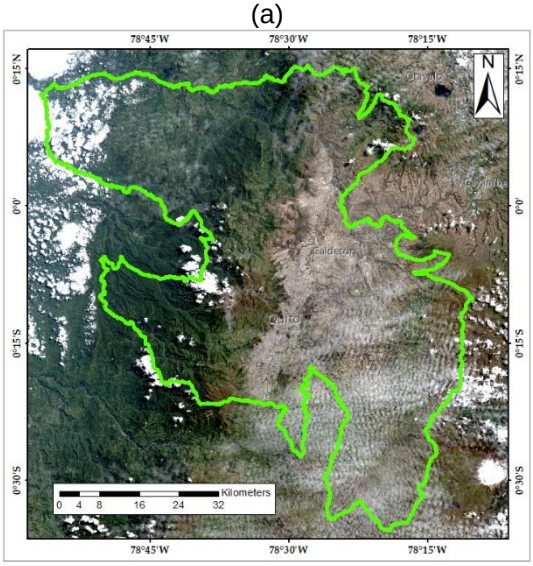
$$DN(u,v) = x_i^f(u,v) + x_i^c(u,v), \quad i = 1,2,3,4,5,6,7, \tag{3.1}$$



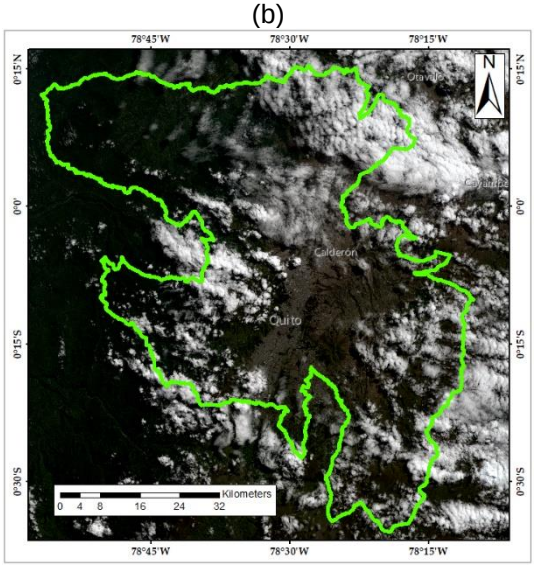
(a)



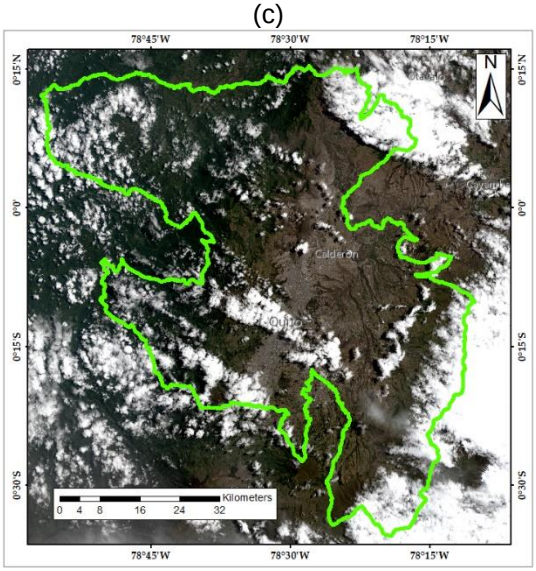
(b)



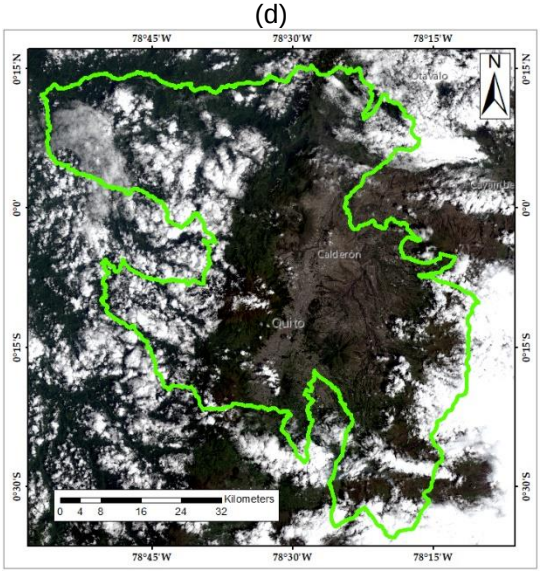
(c)



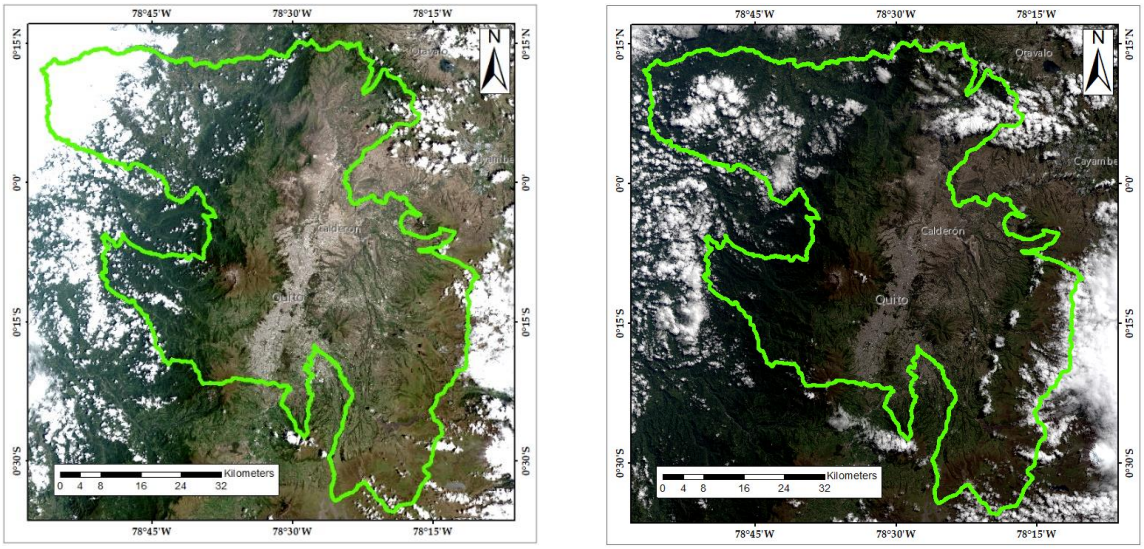
(d)



(e)



(f)



(g) (h)

Figure 3.2. Landsat-8 Images from Quito Metropolitan Area (Path: 10; Row: 60): (a) Image from 2013/10/11; (b) Image from 2013/07/07; (c) Image from 2014/07/26; (d) Image from 2015/07/13; (e) Image from 2015/08/30; (f) Image from 2016/02/06; (g) Image from 2016/10/19; (h) Image from 2013/06/21 (Reference image to ICA evaluation).

where $x_i^f(u,v)$ is the clean cloud-free pixel from each of bands 1–7 and $x_i^c(u,v)$ is the cirrus cloud pixel from each of bands 1–7 obtained with band 9. Equation 3.1 results from the strong linear relationship between the bands found in [116], where $x_i^c(u,v)$ is linearly related to the DN recorded in the cirrus band $c(u,v)$ as follows:

$$x_i^c(u,v) = \propto_i [c(u,v) - min\{c(u,v)\}]. \quad (3.2)$$

The aim is to obtain the slope $\propto_i$ for each band, considering a linear relationship between each multispectral band and band 9 in a homogenous area. Two approaches can be considered to determine this homogenous area. The first approach is a photo-interpretation to find this area by taking, for example, water bodies that have a near-zero pixel value over the near-infrared (NIR) band. However, this approach cannot be used for images that do not contain water bodies. The second approach is to use random areas of a constant size covering the entire region or zones with a specific land use. In this study, we considered the second approach of finding random areas with a size of 10 × 10 km^2, covering the entire study area (Figure 3.3). Smaller regions (250 m * 250 m) were also tested, but the results were identical.

By combining Equation (1) with Equation (2), $x_i^f(u,v)$ can be estimated as follows:

$$x_i^f(u,v) = DN(u,v) - \propto_i [c(u,v) - min\{c(u,v)\}] \quad (3.3)$$

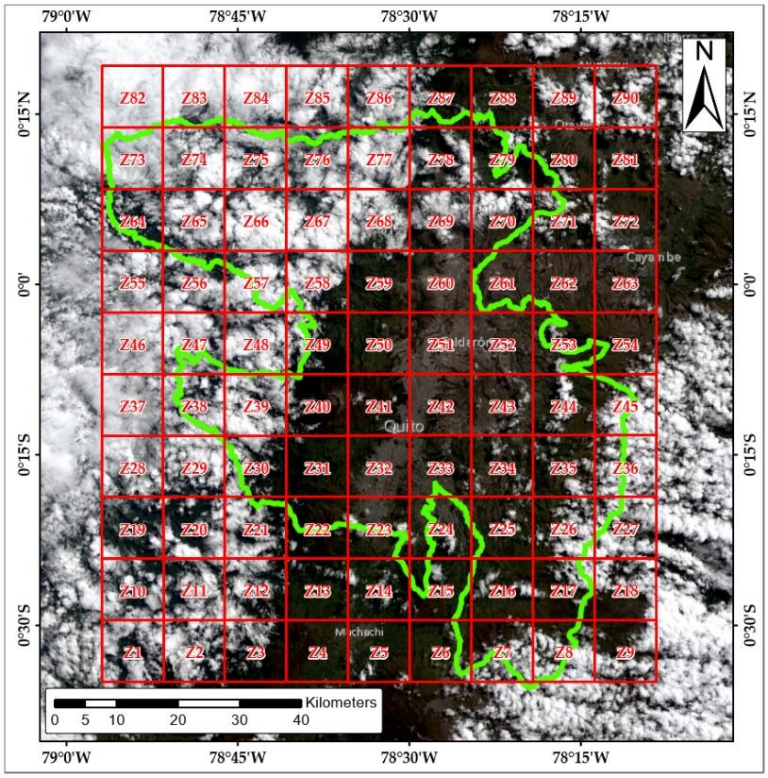


Figure 3.3. Input regions considered to test the ACRM algorithm

3.3.2.2. Independent Component Analysis (ICA)

ICA is a method for finding underlying factors or components from multivariate (multidimensional) statistical data [117]. The relationship is represented as follows:

$$\mathbf{X} = \mathbf{AS} \tag{3.4}$$

where $\mathbf{S}$ is a random vector containing the independent source signal or independent components (IC) with elements $s_1, s_2, \ldots$, and s_n. A is the "mixing" square matrix having elements a_{ij}. $\mathbf{X}$ is the observed signal (mixed) having elements $x_1, x_2, \ldots$, and x_n.

In Equation 3.4, $\mathbf{X}$ represents surface reflectance data from each of bands 1-7 and pixel cirrus data from band 9. The surface reflectance data were obtained by applying atmospheric correction with the fast line-of-sight atmospheric analysis of hypercubes (FLAASH) algorithm [48,118]. FLAASH works as a physical method to obtain surface reflectance, and it allows us to describe the shape of the signatures [49] in ENVI software. The column vector $\boldsymbol{s}$ represents ICs and matrix $\mathbf{A}$ represents the linear transformation. Both $\boldsymbol{s}$ and $\mathbf{A}$ are unknown.

In some studies, ICA is used to separate some parts of satellite images by considering their bands as ICs. The algorithm achieves cloud removal by considering that each IC is a linear mixture of bands 1–7 and 9. Band 9 is used to delineate the cloud component in the IC [102,103].

ICA works with a non-Gaussian distribution, where ICs (surface reflectance and pixel cloud data) are not normally distributed, because various surface types and cloud types produce different reflectance values. The robust FastICA algorithm can be applied to estimate an unmixing matrix $\mathbf{W}$, which is the inverse of mixing matrix $\mathbf{A}$ [110]. The source vector $\boldsymbol{s}$ can be obtained by inverting Equation 3.4 as follows:

$$\boldsymbol{s} = \mathbf{A}^{-1}\mathbf{X}. \tag{3.5}$$

Band 9 (cirrus band, which is a part of $\mathbf{X}$) is considered the sum of eight products (bands 1–7 and 9) for each IC: the product of each source vector with its coefficients in $\mathbf{A}$. Equation 3.6, derived from Equation 3.4, allows us to obtain the cloud pixel value $\mathbf{x_{17}}$ as follows:

$$\mathbf{x_{1-7}} = \mathrm{a}_{1-7}\boldsymbol{s_c}, \tag{3.6}$$

Where a_{1-7} is the coefficient of $\boldsymbol{s_c}$ in matrix $\mathbf{A}$ corresponding to the reflectance data of bands 1-7. The largest factor in the row corresponding to band 9 of $\mathbf{A}$ determines the $\boldsymbol{s_c}$ to be used to obtain the cloud reflectance data $\boldsymbol{x_c}$. The final reflectance-free data x_f is obtained by subtracting the original reflectance data from each band $\boldsymbol{x_o}$ by the cloud reflectance data from each band $\boldsymbol{x_c}$ (Equation 3.7).

$$x_f = x_o - x_c. \tag{3.7}$$

3.3.2.3. Normalized Difference Vegetation Index (NDVI)

NDVI is an index that allows to obtain information about the greenest vegetation considering red and NIR bands of a sensor [54]. In the case of Landsat-8 OLI, NDVI is calculated using bands 4 (red band) and 5 (NIR band). The NDVI in a Landsat-8 OLI image is computed as follows (Equation 3.8):

$$NDVI = \frac{B5 - B4}{B5 + B4} \tag{3.8}$$

NDVI is one of the most commonly used remote sensing vegetation indices [119,120], and it is considered an environmental index owing to its strong relationship with the land surface (e.g., surface temperature, vegetation cover, land use) and meteorological data (e.g., temperature, humidity) [121]. Moreover, NDVI is used to validate and compare results between sensors by considering future environmental applications [122].

3.3.2.4. Evaluation and Validation

In order to validate the efficiency of ACRM and ICA cloud removal methods in the computation of environmental indices, the NDVI was computed in the original Landsat-8 images after applying both algorithms. Then, the images were compared with a MODIS NDVI product resampled to a spatial resolution of 30 m, assuming a similar period of

Landsat-8 data used. A MOD13Q1 product (NDVI 16-Day L3 Global 250 m version 6) was used as reference data, considering that MODIS is a ready-to-use product [9,56] and is evaluated in vegetation phenology. The validation was tested in a small area where cirrus clouds are present, which allowed us to evaluate the performance of the algorithms to remove clouds and to estimate environmental indices. The methodology adopted in this work is presented in the flowchart shown in Figure 3.4.

3.4 Results

3.4.1. Cloud Removal Using ACRM

The ACRM algorithm was applied to ten images considered in this study. The code was programmed in R Studio with the raster package. The main objective was to obtain the best correlation (R^2) between bands 1–7 and band 9 in selected areas of the images with cirrus clouds.

The first step was to choose the zones to evaluate the algorithm in a geographic information system (GIS) covering the entire study area in Quito. These areas, called zones (Z), are 10 km × 10 km regular grids covering the study area (Figure 3.3). Subsequently, the algorithm was applied, and the best-fit regions with the best R2 coefficients between each multispectral band (1-7) and band 9 (Table 3.2) were evaluated.

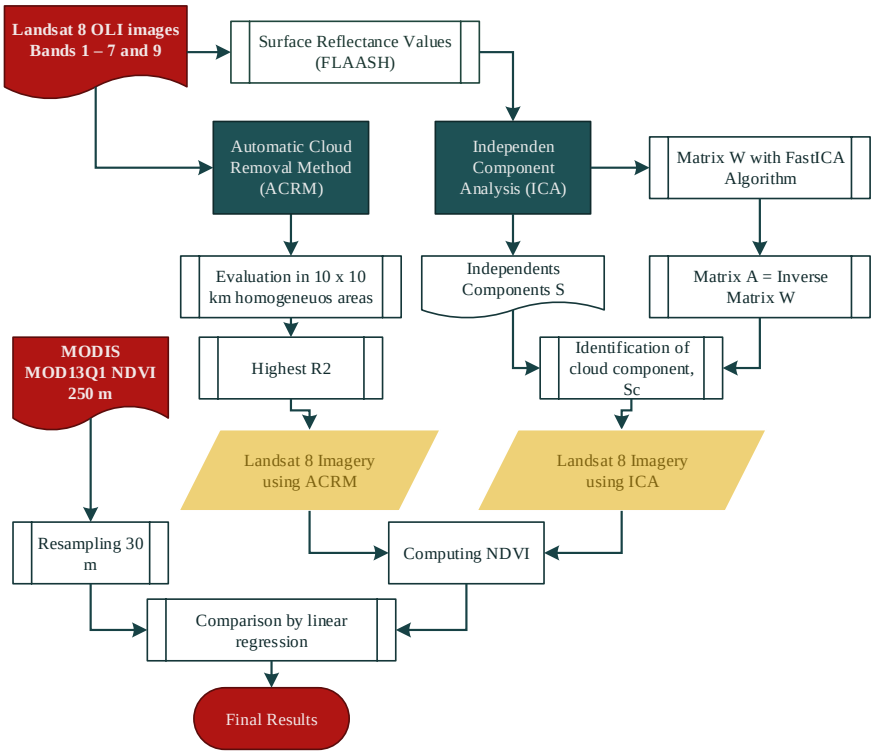
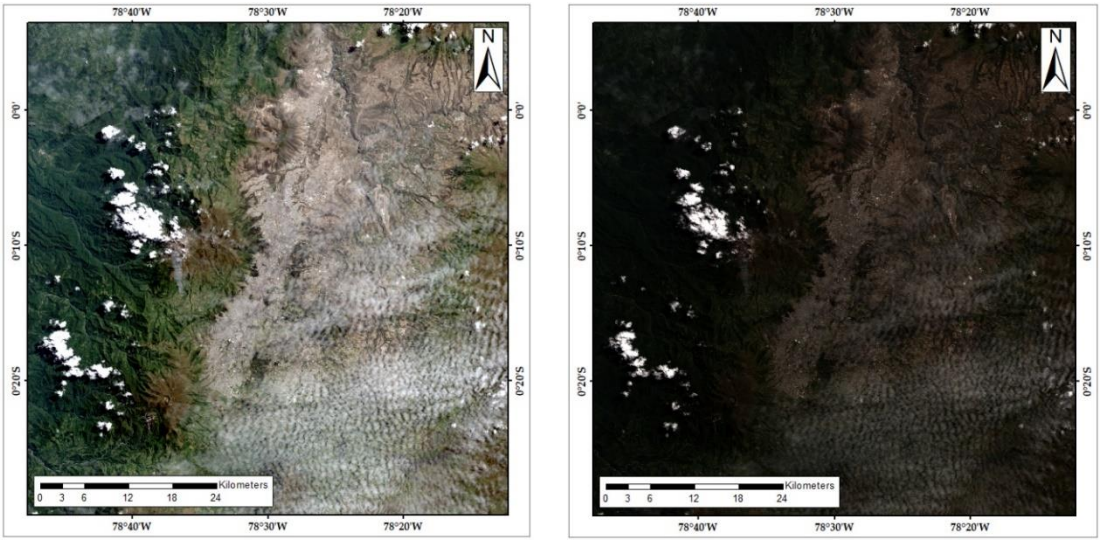


Figure 3.4. Flowchart of the methodology adopted to perform a comparison between ACRM and ICA algorithms.

Table 3.2 lists the highest R^2 coefficients obtained in the application of the algorithm, considering only values higher than 0.85. Slope values are lower than 0.18. These results are shown in Figure 3.5 (see Section 3.3.4.3).

ACRM was also tested considering an image from Pedernales and an image from Sydney (Table 3.3). In Pedernales, the R^2 coefficients had values lower than 0.68. Better results were obtained over Sydney with higher R^2 coefficients (higher than 0.97). To corroborate the results of R^2 coefficients (Figure 3.6), we confirmed that the image of Pedernales is practically unchanged by the algorithm, while the algorithm removes all the clouds in the image of Sydney.



(a) (b)

Figure 3.5. Landsat-8 Images from Quito Metropolitan Area (Path: 10 Row: 60): Image from 2014/07/26 (a) Original Image; (b) Image applied ACRM

Table 3.2. Linear regression results between bands 1–7 and 9 in the Quito study area for different dates.

Band	Quito (11/10/2013)		Quito (07/26/2014)		Quito (07/13/2015)		Quito (02/06/2016)	
	R^2	Slope (α)	R^2	Slope (α)	R^2	Slope (α)	R^2	Slope (α)
B2	0.96	0.05	0.93	0.02	0.95	0.03	0.95	0.03
B3	0.96	0.05	0.93	0.02	0.95	0.03	0.95	0.03
B4	0.96	0.05	0.93	0.02	0.95	0.02	0.95	0.02
B5	0.88	0.02	0.85	0.01	0.91	0.02	0.85	0.01
B6	0.85	0.02	0.89	0.17	0.88	0.02	0.89	0.03
B7	0.86	0.02	0.88	0.02	0.87	0.02	0.88	0.02
Band	**Quito (07/07/2013)**		**Quito (08/30/2015)**		**Quito (10/19/2016)**		**Quito (21/06/2013)**	
	R^2	Slope (α)	R^2	Slope (α)	R^2	Slope (α)	R^2	Slope (α)
B2	0.96	0.05	0.93	0.02	0.97	0.03	0.95	0.03
B3	0.96	0.06	0.93	0.02	0.97	0.03	0.95	0.03
B4	0.95	0.05	0.93	0.02	0.97	0.02	0.95	0.02
B5	0.85	0.03	0.85	0.01	0.95	0.02	0.85	0.01
B6	0.90	0.06	0.89	0.17	0.92	0.02	0.89	0.03
B7	0.89	0.06	0.88	0.02	0.89	0.03	0.88	0.02

Table 3.3 Linear regression results between bands 1–7 and 9 in the other evaluated zones.

Band	Sydney (2013/10/04)		Pedernales (2016/05/13)	
	R^2	Slope (α)	R^2	Slope (α)
B2	0.97	1.70	0.67	0.69
B3	0.99	1.63	0.68	0.68
B4	0.98	1.68	0.67	0.62
B5	0.98	1.74	0.67	0.52
B6	0.99	1.11	0.63	0.44
B7	0.98	1.02	0.53	0.58

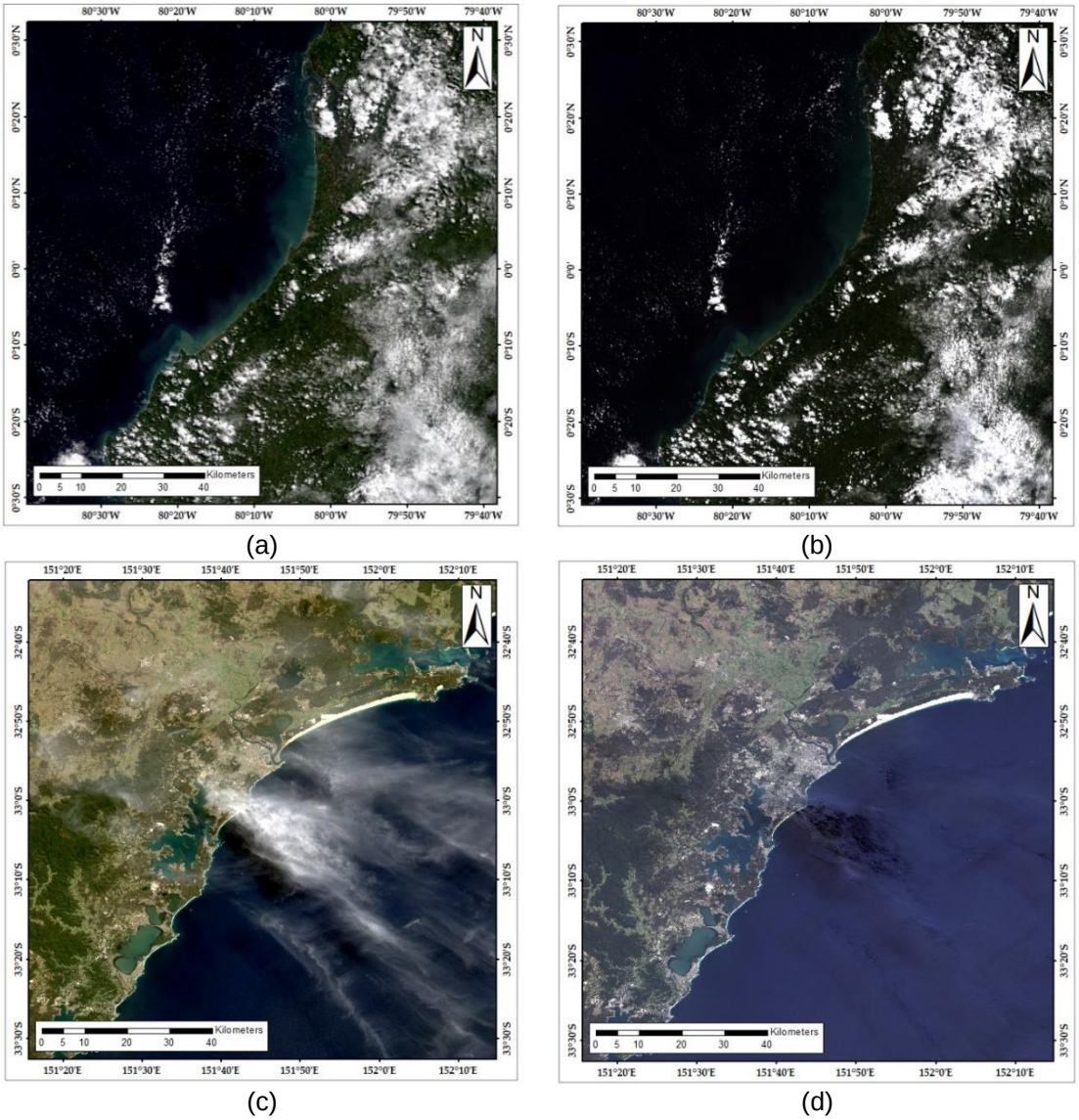


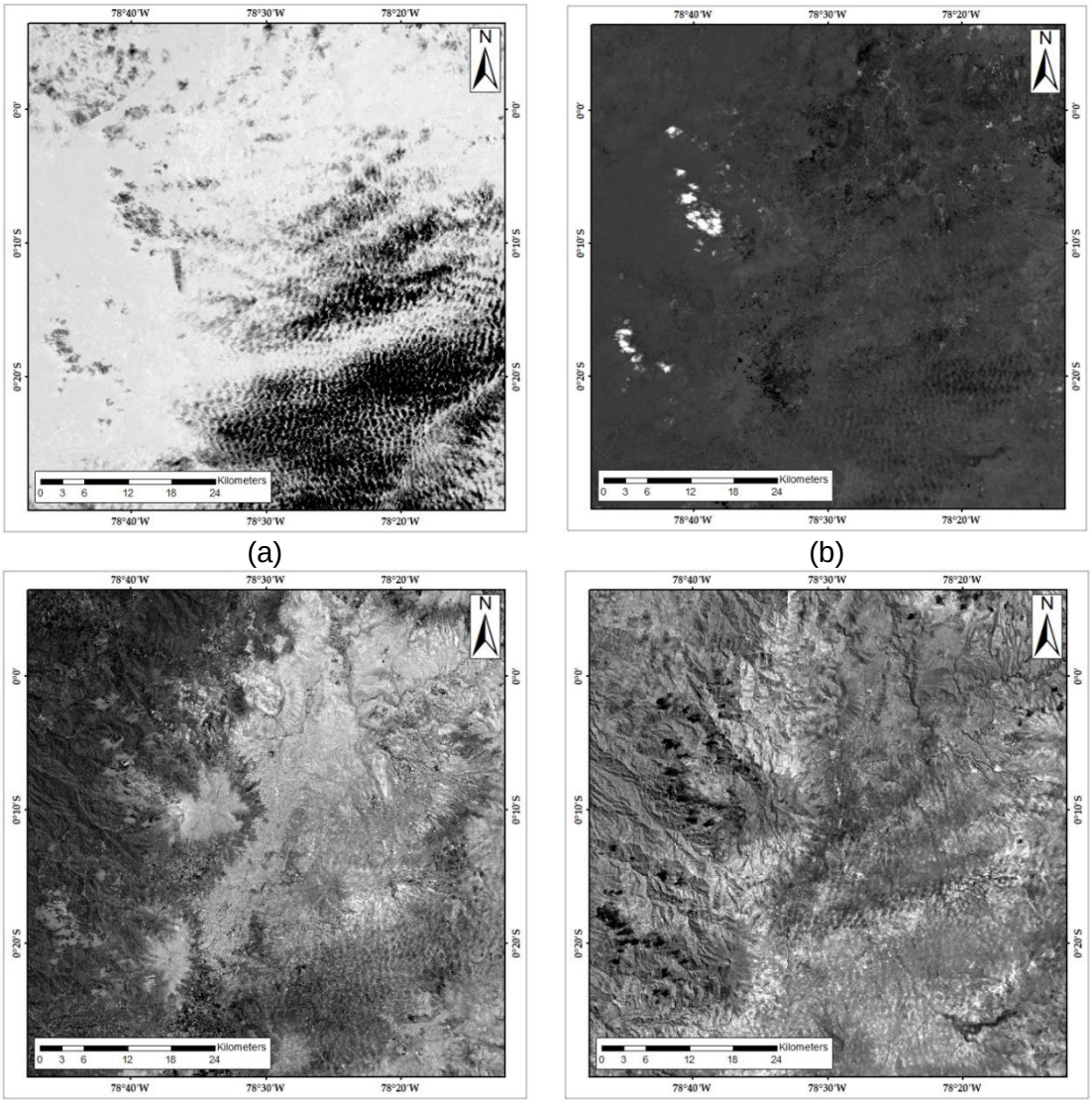
Figure 3.6. Landsat 8 OLI images (a) Original image from Pedernales; (b) Image after applied ACRM in Pedernales; (c) Original image from Sidney; (d) Image after applied ACRM in Sidney

3.4.2. Cloud Removal Considering ICA

The ICA algorithm was applied only to the Quito image from 26/07/2014, which shows clouds over the study area. Different software were used (R Studio, ENVI, ERDAS) to obtain the different parameters showed in the Equation 3.4. The principal inputs to the algorithm were the surface reflectance data of multispectral bands (calculated with

FLAASH correction from ENVI) and the DN from band 9. Furthermore, the IC for the selected image was obtained in ENVI software with the FastICA algorithm [110] (Figure 3.7). The matrix A from Equation 3.6 was obtained using the ICA algorithm in ERDAS software (Table 3.4), and s_c was selected as s_6, which had the high absolute value of 4.011×10-2 in the row of band 9. Then, to obtain the input data for Equation 3.7, the product of the coefficient in the column for each band at s_6 with each IC was used. The results are shown in Figure 3.8. Again, as in ACRM, the result was not satisfactory in comparison with the original image (see Section 3.3.4.3).

Moreover, to corroborate that the application of the ICA algorithm does not provide satisfactory results for Quito, some scatterplots were computed with respect to a cloud-free reference image (Figure 3.9). The scatterplots show a linear correlation between the reference image (Figure 3.2h) and the images with and without ICA correction (Table 3.5), which indicates that the ICA algorithm does not work properly for Quito.



(a) (b)

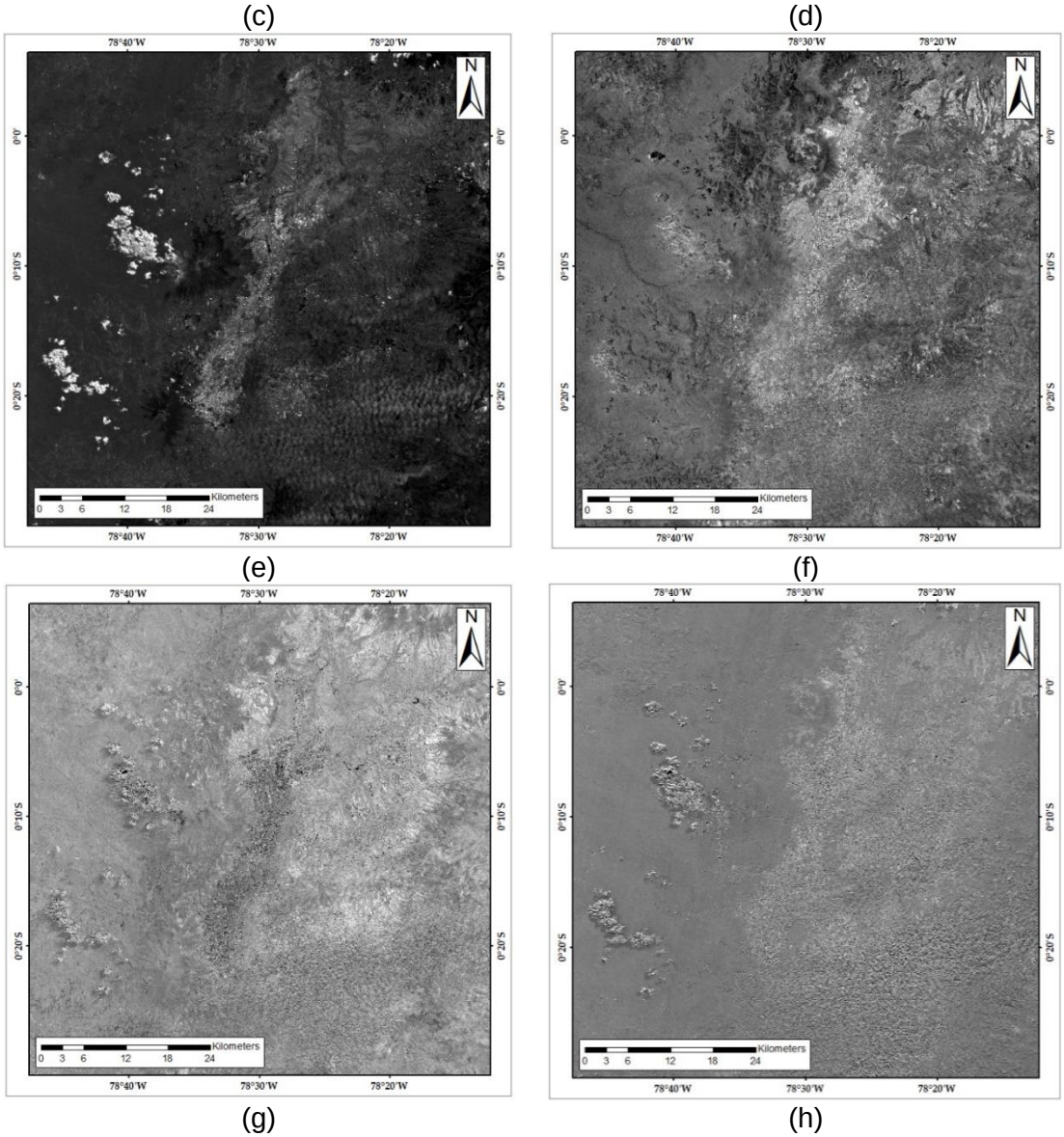
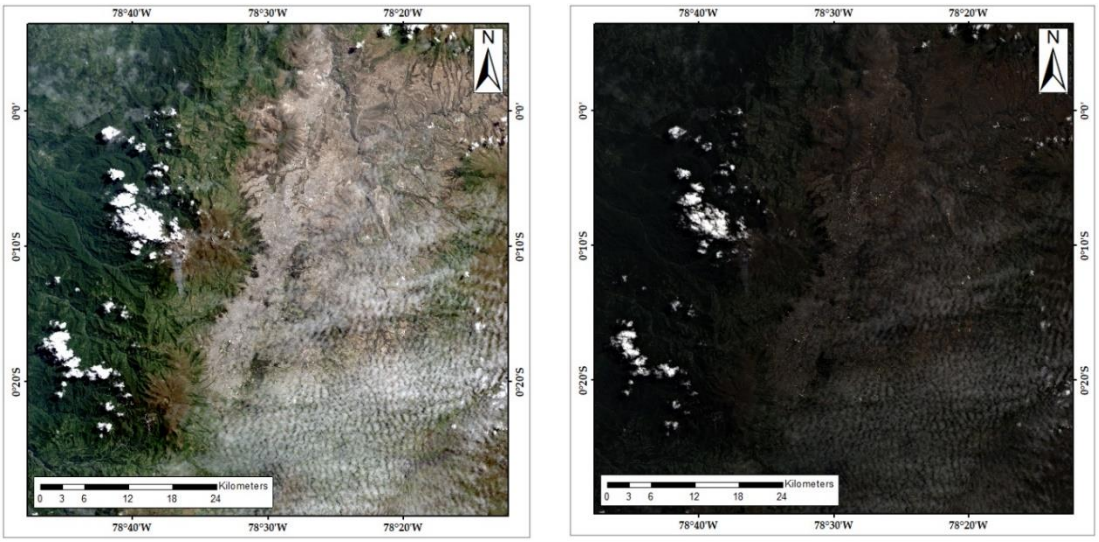


Figure 3.7. (a–h) are first, second, …, and eighth independent components, respectively.

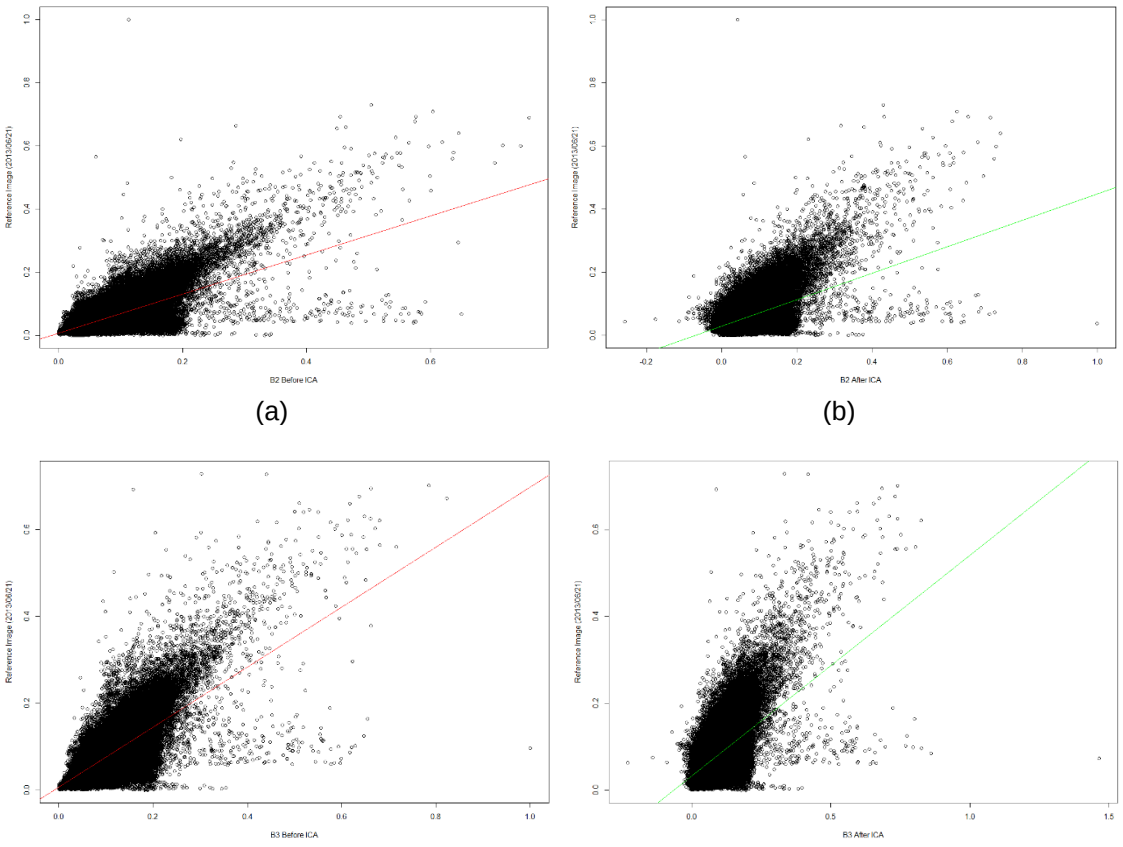
Table 3.4. Coefficients (×10−2) of A

Band	S1	S2	S3	S4	S5	S6	S7	S8
B1	4.719	0.678	0.653	9.672	1.731	1.818	1.308	0.207
B2	4.613	0.939	0.494	9.192	1.628	1.661	1.722	0.372
B3	4.537	0.802	1.153	8.826	1.645	1.644	2.201	1.149
B4	4.487	0.696	0.851	8.954	1.493	1.692	3.413	1.006
B5	2.824	0.475	0.524	6.962	1.743	1.148	-1.815	7.568
B6	0.236	0.764	1.266	7.093	1.508	1.632	3.497	3.671
B7	0.256	0.901	1.214	6.417	-0.022	1.794	3.746	1.656
B9	-0.021	-0.023	0.018	-0.152	0.984	4.011	0.617	0.108



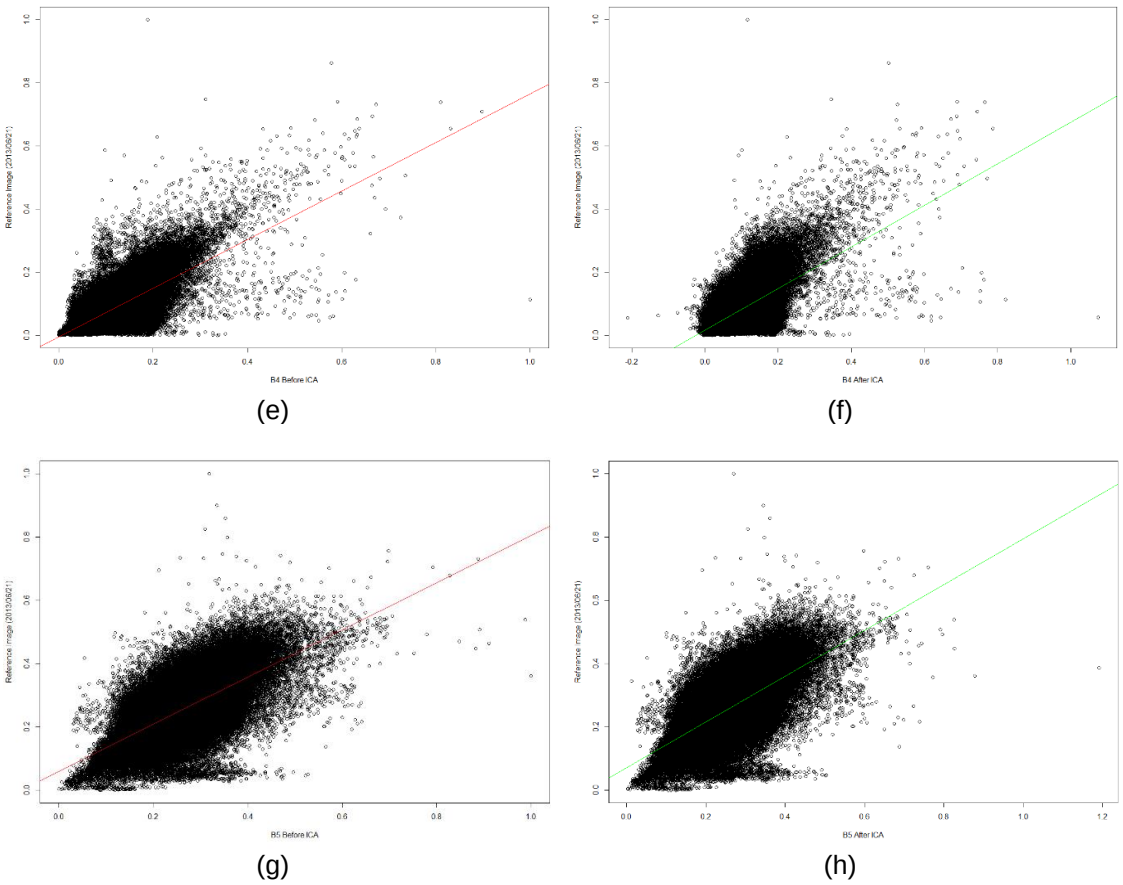
(a) (b)

Figure 3.8. Landsat-8 Images from Quito Metropolitan Area (Path: 10 Row: 60): Image from 2014/07/26 (a) Original Image; (b) Image after applied ICA.



(a) (b)

(c) (d)



(e) (f)

(g) (h)

Figure 3.9. Scatterplots of bands 2-5. (a, c, e, g) Left an image before ICA algorithm implementation vs. reference image. (b, d, f, h) Right image considers ICA algorithm implementation vs. Reference image. Reference image is from June 21, 2013 to evaluate ICA (Figure 3.2h).

As indicated in Table 3.5, if ICA is applied, the algorithm changes the surface reflectance values; in comparison with a cloud-free image, the correlation decreases.

Table 3.5. Linear Regression. R^2 coefficients before and after ICA computation

Band	R^2 before	R^2 after
B2	0.43	0.20
B3	0.49	0.26
B4	0.53	0.33
B5	0.49	0.47

3.4.3. Validation – NDVI Computation

As mentioned previously, one of the main objectives of the cloud removal in high-altitude areas is to obtain a better accuracy in the computation of environmental indices, such as NDVI. Therefore, in the process of validation of the proposed algorithms, the NDVI values for a selected area (Quito airport) with a high density of cirrus clouds were computed (Figure 3.10).

NDVI values were compared to the MODIS MOD13Q1 product and resampled to a spatial resolution of 30 m to enable them to be related to Landsat data. The MODIS product is of a nearer date (07/28/2014) to the Landsat-8 image (Figure 3.11a). The

validation compares the reference NDVI product (MODIS MOD13Q1 resampled) and the NDVI computed through the Landsat-8 image. NDVI values are computed considering the original surface reflectance of the Landsat-8 image (Figure 3.11b) and the surface reflectance of the images after applying the two algorithms for removing cirrus clouds: i) ACRM (Figure 11c) and ii) ICA (Figure 11d).

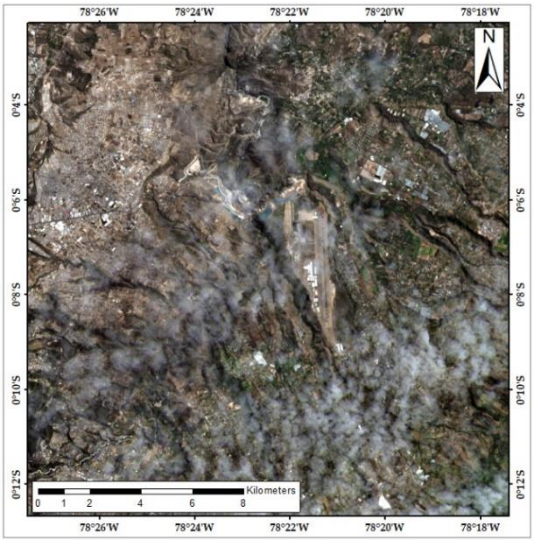
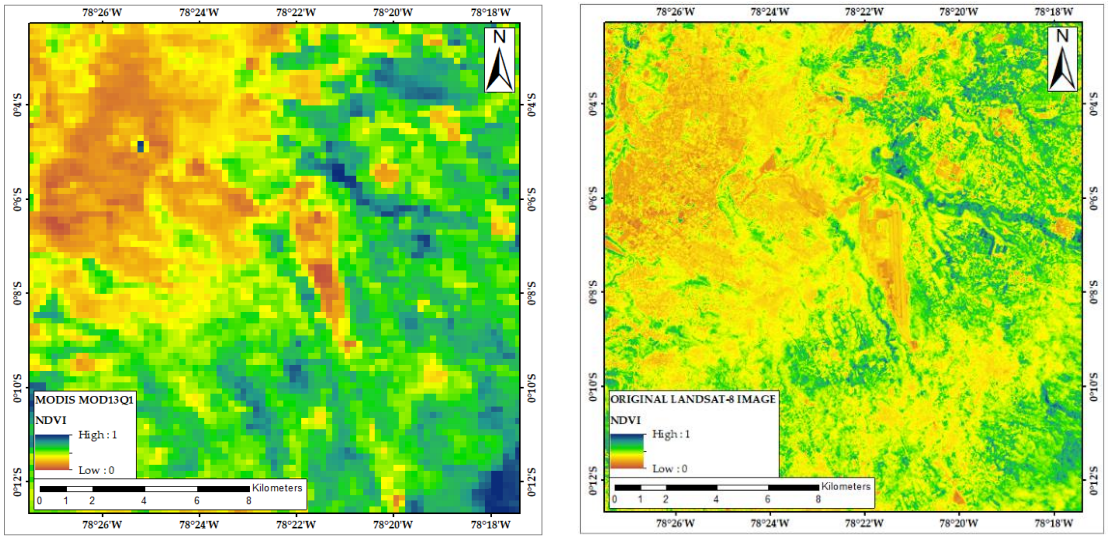
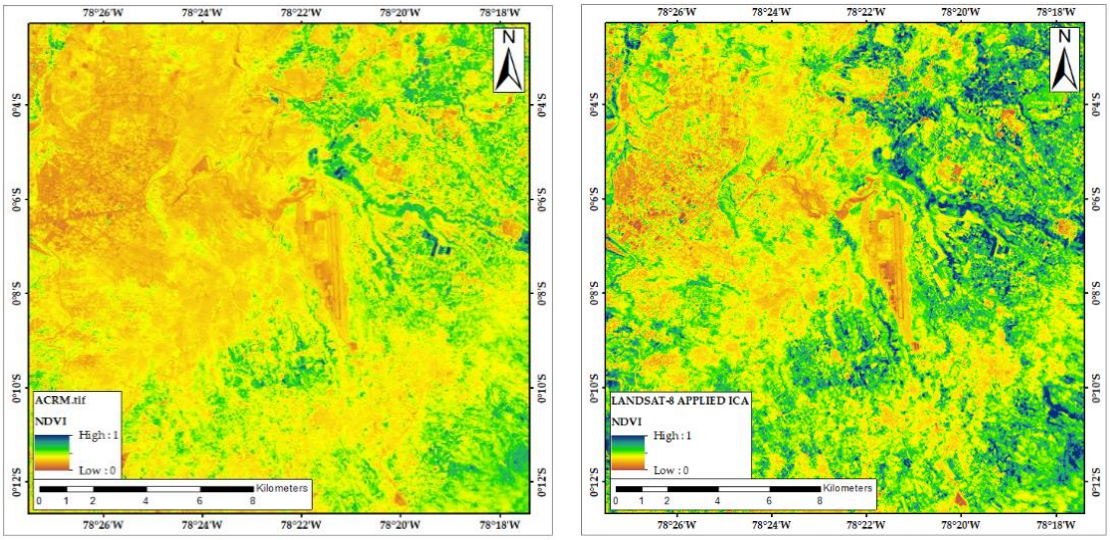


Figure 3.10. Area evaluated in Quito airport to compute NDVI (Landsat-8 image from 07/26/2014).



(a) (b)



(c) (d)

Figure 3.11. NDVI computed from (a) MODIS NDVI 30 m resampled image; (b) original Landsat-8 image; (c) Landsat-8 image after cloud correction using the ACRM algorithm; (d) Landsat-8 image after cloud correction using the ICA algorithm.

In order to compare MODIS NDVI and the other NDVI computations, a linear regression was established to obtain R^2 coefficients, and the results showed that the highest R^2 (0.426) is obtained after applying ACRM. On the other hand, the lowest coefficient is obtained after applying ICA with an R^2 value of 0.262 (Table 3.6).

Table 3.6. Linear Regression between MODIS NDVI and NDVI computed from each cloud removal method.

NDVI Computation with	**R^2**
Original Image with Surface Reflectance Data	0.396
After ACRM algorithm	0.428
After ICA algorithm	0.262

3.4.4. Improvement of ACRM

According to the preliminary results (Table 3.6), the ACRM algorithm yielded the highest R^2 to calculate environmental indices; nevertheless, one improvement of the ACRM method was developed to remove clouds in Landsat-8 OLI images of high-elevation areas [35]. This development attempts to find the best-fit slope in the ACRM algorithm, established in Equation 3.3, to remove clouds in order to compute environmental indices. When ACRM was applied to an image of Quito, the slope parameter presented low values, which led us to conclude that the correction to remove clouds does not work properly when it takes values close to 0 (Table 3.2).

A previous work used a fixed slope value [32]. The main improvement in the ACRM algorithm was to find the highest R^2 coefficients in the homogeneous zones and the best-

fit slope to remove clouds. Several slope values from 0 to 100 (in increments of 0.1) were tested. Therefore, the improvement was to find the highest R^2 with the fittest slope testing several slopes values. This procedure was implemented in R Studio software.

To compare and validate the best-fit slope, NDVI was computed for the original image (07/26/2014) after applying the ACRM algorithm and compared with the MODIS NDVI, resulting in the highest R^2 (0.5077) with a slope value of 2.9 (Figure 3.12).

The slope value of 2.9 allowed to a visualization without clouds (Figure 3.13 and Figure 3.14). However, this value is not necessarily the same in each case. The slope value must be investigated for each case, in order to find the best fit to the corresponding area and image.

The results of comparing the R^2 between the different methods are shown in Figure 3.15. The improved ACRM shows the highest R^2 value (0.5077), and visually, it removes clouds to yield a clean image (Figure 13d). Thus, the improved ACRM works satisfactorily over the study area.

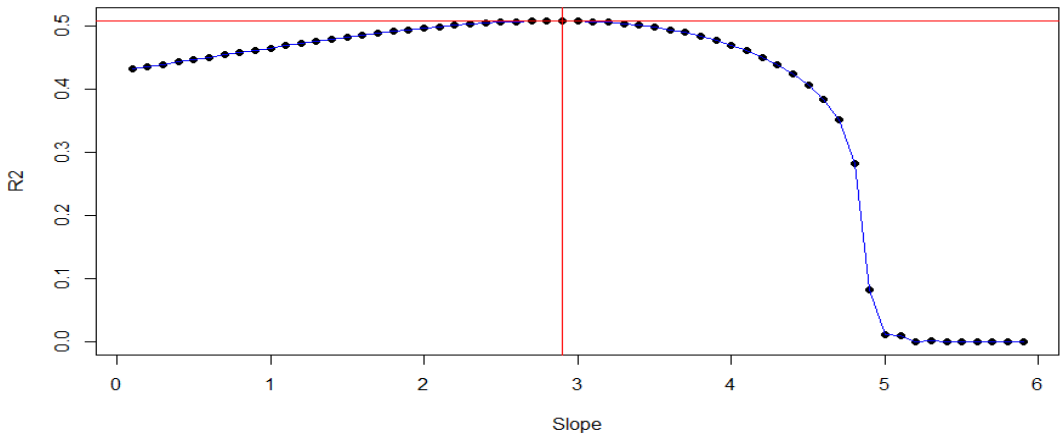
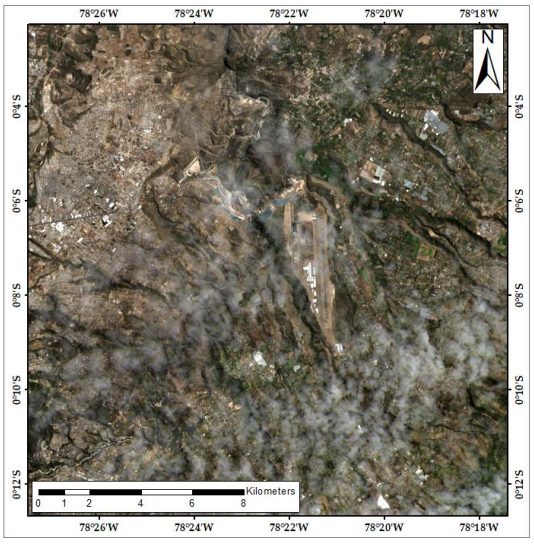
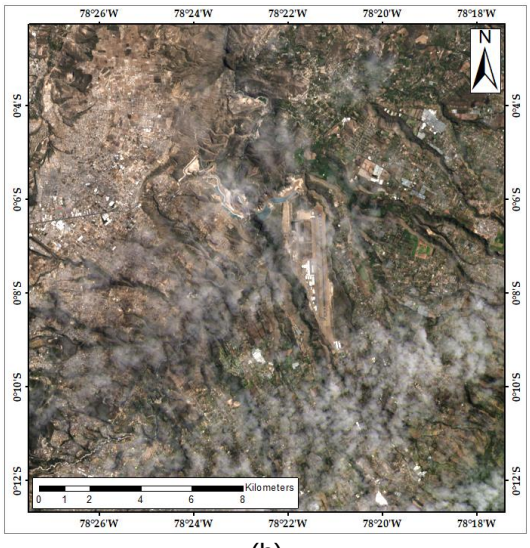


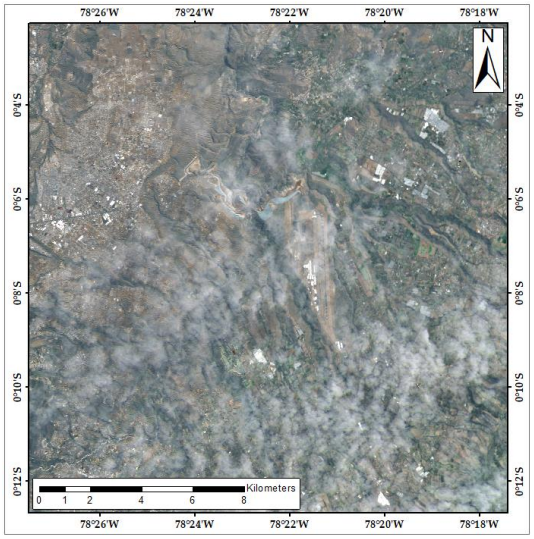
Figure 3.12. Comparison between NDVI obtained using ACRM for each slope tested (dots) with the MODIS NDVI. The red lines indicate the highest R2 and the corresponding slope.



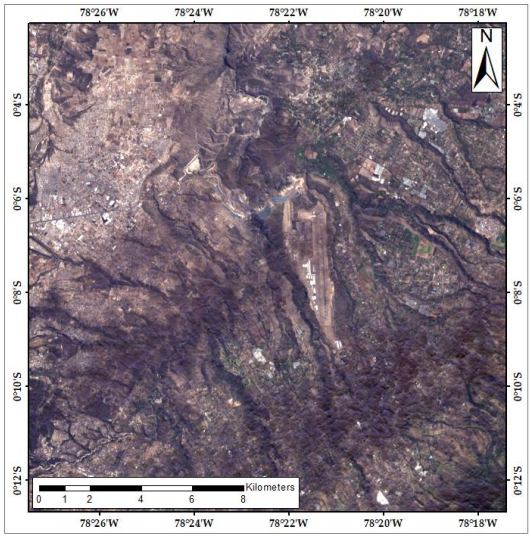
(a)



(b)

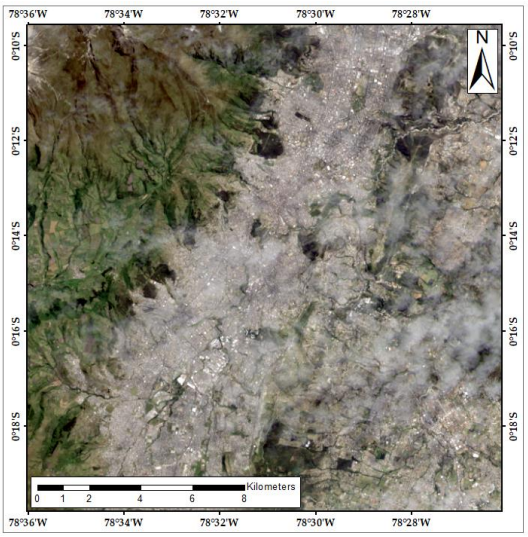


(c)

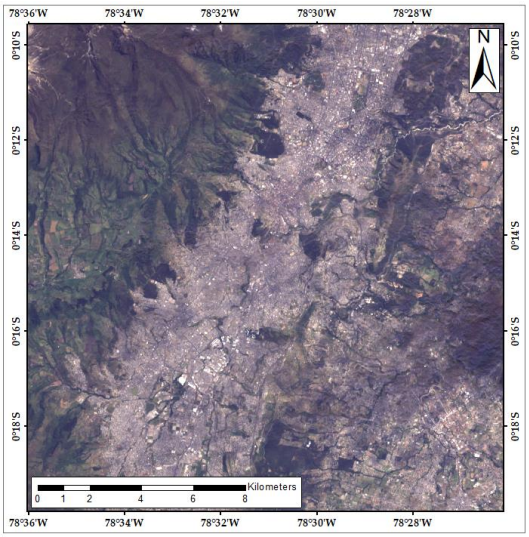


(d)

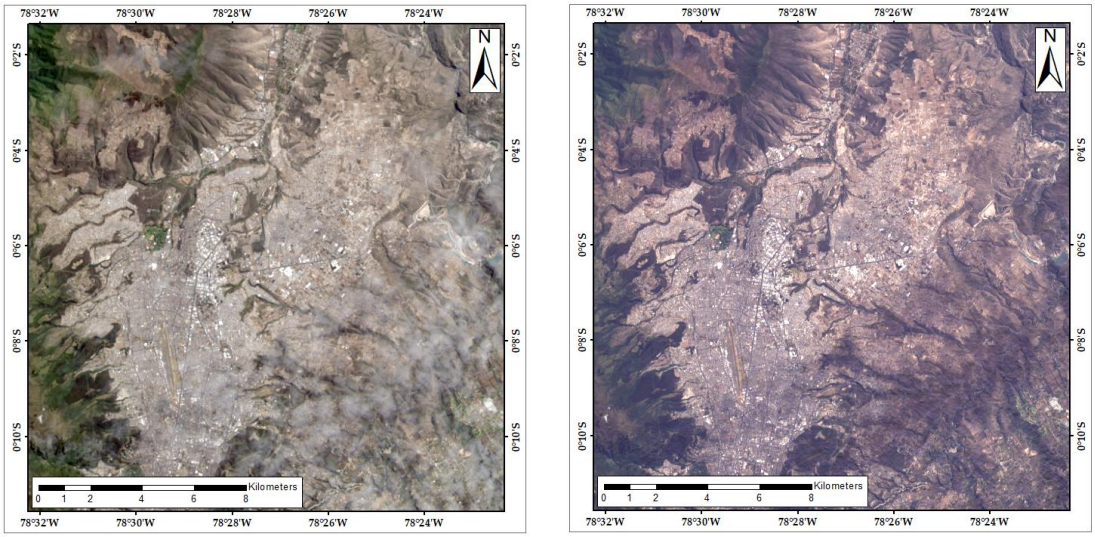
Figure 3.13. Images of Quito airport used to compute NDVI (based on Landsat-8 image from 07/26/2014) (a) original image; (b) image obtained after applying the ACRM algorithm; (c) image obtained after applying the ICA algorithm; (d) image obtained after applying the improved ACRM algorithm.



(a)



(b)



(c) (d)

Figure 3.14. Comparison of result applying the ACRM improvement (b),(d) in different regions vs. the surface reflectance image (a),(c).

In order to validate the ACRM, a new image (11/10/2013) with similar properties was used in the same area. The results show a higher R^2 (0.5283) with a slope value of 2.8 in the ACRM (Figure 3.16).

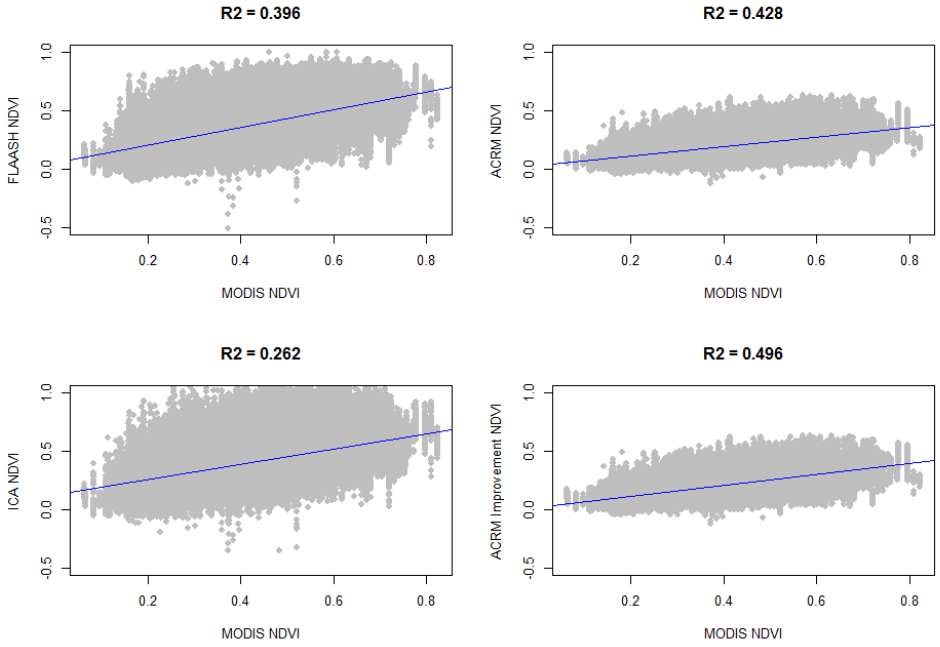


Figure 3.15. Comparison between MODIS MOD13Q1 and the different NDVI values obtained from the application of the different algorithms in the Landsat-8 image (07/26/2014) for removing clouds.

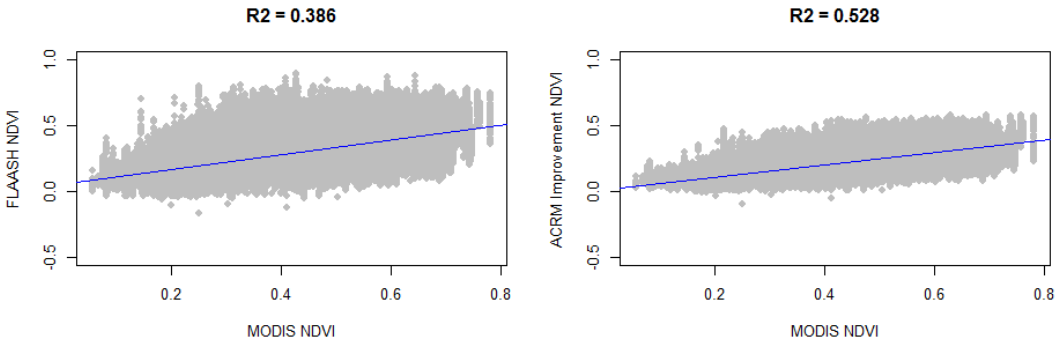


Figure 3.16. Comparison between MODIS MOD13Q1 and the NDVI value obtained from Original Surface Reflectance data (FLAASH Correction applied) and ACRM improved in the Landsat-8 image (11/10/2013) for removing cirrus clouds.

3.5 Discussion and Conclusion

Two algorithms, ACRM and ICA, were employed to remove cirrus clouds in Landsat-8 images with the cirrus band (B9) [10], in Quito city. The main advantage of these two methods is that they do not use additional images to patch data, in contrast to other methods [99,100,107,108]. These methods use the same image to remove thin cloud without the insertion of pixel values from other images. In this work, because cirrus clouds could have a great impact in the computation of environmental indices such as NDVI, these two methods were tested and compared with the aim of evaluating their applicability to accurately compute NDVI for an area located in the Andean region.

ACRM generated satisfactory results for images with conditions similar to Sydney [35]. The same original image of Sydney was used to reproduce the correct application of ACRM, which yielded an R^2 coefficient higher than 0.95, with slopes higher than 1. These satisfactory results were also evident from visual inspection, because clouds were adequately removed (Figure 3.6d). When the ACRM algorithm was tested for images of Quito from different dates, the results showed R^2 coefficients higher than 0.90 in most of the cases but with low slope values (lower than 0.1 in most of the cases for all bands) (Table 3.3). The low slope values indicate poor correction. Moreover, it is evident from visual inspection that this algorithm does not remove the cirrus clouds over the images (Figure 3.5). Another area, Pedernales, was chosen to test the algorithm because it has similar characteristics to Sydney. The results for this area are also unsatisfactory for the clouds removal (Figure 3.6b).

The other algorithm tested to remove cirrus clouds was ICA [102], which is a blind source method that attempts to obtain the cloud component of images [110]. All ICs contain free pixel data and cloud noise, and the noise should be removed, considering all image data to have a non-Gaussian distribution [117]. ICA was tested for images of Quito, and the results were compared with a cloud-free image (image with surface reflectance data). The results are unsatisfactory because the correlation was worse than the case without

applying ICA (Table 3.4). For example, in band 4, the R^2 value obtained in comparison with the cloud-free image was 0.33; the value without applying ICA was 0.53.

In order to validate the results, NDVI was computed. In the first approximation, the results were compared with a reference image product (MODIS MOD13Q1). The results showed the highest R^2 when the ACRM algorithm was applied; these values were higher than those obtained with ICA or those of the surface reflectance data. Finally, an improvement to ACRM was proposed. This algorithm had two main objectives: (i) visually remove clouds and (ii) improve the pixel values to compute environmental indices. The ACRM algorithm was improved, so that the homogeneous area has the highest R^2 coefficient value and the slope should be significant to reduce the density of cirrus clouds. In the case of the study area (Quito), the first condition was achieved with a high R^2 coefficient between Landsat multispectral bands and band 9 in a homogeneous area (Table 3.1). The challenge was to achieve cloud correction using ACRM. Therefore, we tested different slope values [32] between 0 and 100, and the best-fit slope value of 2.9 was obtained. This approach proved to be a good alternative to the previous algorithms tested (Figure 3.13). In order to validate this new approach, the NDVI values were computed and compared with the reference NDVI values (MODIS). This new approach yielded higher R^2 values (Figure 3.15 and Figure 3.16). The ACRM Improved using the highest R^2 value can approximate to other products ready to use like MODIS NDVI, finding a better relationship than other algorithms or methods, and a considerable best performance, since can be applied to Landsat 8 data, which have a spatial resolution of 30 m.

The preliminary results show that the algorithms to remove cirrus clouds (ACRM and ICA) do not work properly in the geographical conditions considered in this study, leading us to suppose that there are other factors such as altitude and closeness to the equator that influence the results. Therefore, future research should focus on testing these algorithms in different regions around the world to determine the best method for each area or to identify better alternatives to improve the cloud removal algorithms. Moreover, in some parts of the world such as Quito, Landsat images are affected by a high cloud density throughout the year, limiting the time frame to obtain phenology data at a spatial resolution of 30 m. Nevertheless, the ACRM improved can help in a more accurate computation of environmental indexes when compared to other algorithms or methods.

4. Article 2: Assessment of remote sensing data to model PM10 estimation in cities with a low number of air quality stations. A case of study in Quito, Ecuador.

Cesar I. Alvarez-Mendoza[1,2*], Ana Teodoro[1,3], Nelly Torres[2] and Valeria Vivanco[2]

[1] University of Porto, Department of Geosciencies, Environment and Land Planning, Faculty of Sciences, Rua Campo Alegre 687, Porto 4169-007, Portugal; calvarezm@ups.edu.ec

[2] Universidad Politécnica Salesiana, Grupo de Investigación Ambiental en el Desarrollo Sustentable, Carrera de Ingeniería Ambiental, Quito, Ecuador; lramirez@ups.edu.ec

[3] Earth Sciences Institute (ICT), Pole of the FCUP, University of Porto, Porto, Portugal; amteodor@fc.up.pt

Academic Editors: Domenico M. Cavallo, Andrea Spinazzè

Received: 14 June 2019 / Accepted: 19 July 2019 / Published: 21 July 2019

Journal: **Environments. MDPI. Special Issue Air Quality Assessment for Environmental Policy Support: Sources, Emissions, Exposures and Health Impacts**

Environments **2019**, *6*(7), 85; https://doi.org/10.3390/environments6070085

4.1 Abstract

The monitoring of air pollutant concentration within cities is crucial for environment management and public health policies in order to promote sustainable cities. In this study, we present an approach to estimate the concentration of particulate matter of less than 10 μm diameter (PM10) using an empirical land use regression (LUR) model and considering different remote sensing data as the input. The study area is Quito, the capital of Ecuador, and the data were collected between 2013 and 2017. The model predictors are the surface reflectance bands (visible and infrared) of Landsat-7 ETM+, Landsat-8 OLI/TIRS and Aqua-Terra/MODIS sensors and some environmental indexes (Normalized Difference Vegetation Index – NDVI; Normalized Difference Soil Index - NDSI, Soil-Adjusted Vegetation Index – SAVI; Normalized Difference Water Index - NDWI and Land Surface Temperature (LST). The dependent variable is PM10 ground measurements. Furthermore, this study also aims to compare three different sources of remote sensing data (Landsat-7 ETM+, Landsat-8 OLI and Aqua-Terra/MODIS) to estimate the PM10 concentration, and three different predictive techniques (stepwise regression, partial least square regression and artificial neuronal network (ANN)) to build the model. The models obtained are able to estimate PM10 in regions where air data acquisition is limited or even does not exist. The best model is the one built with an ANN, where the coefficient of determination (R^2 = 0.68) is the highest and the root-mean-square error (RMSE = 6.22) is the lowest among all the models. Thus, the selected model allows the generation of PM10 concentration maps from public remote sensing data, constituting an alternative over other techniques to estimate pollutants, especially when few air quality ground stations are available.

Keywords: Remote Sensing, air quality modeling, air quality monitoring, PM10, LUR

4.2 Introduction

Due to some factors as air pollutants permanency over the time, the air quality has decreased in recent years, all over the world. One of the direct indicators of air quality is particulate matter with an aerodynamic diameter lower than 10 μm, usually called PM10 [123]. It is well-known that PM10 has a negative environmental impact on outdoor air quality and that it that is linked to public health problems such as cardiovascular and respiratory diseases [124,125]. Many cities around the world are monitoring PM10 in order to prevent environmental problems. However, this monitoring process needs to be improved in order to establish reliable environmental policies [126]. Thus, understanding

the spatial distribution of PM10 requires a scientific and accurate basis to locate the possible sources of this pollutant in cities, in order to avoid environmental problems linked to air quality.

The air quality monitoring network (AQMN) is a classical procedure to monitor PM10 in cities. However, some difficulties are found, for instance, high maintenance cost by station [66], a low quantity of stations in large cities or non-representative spatial distribution [67]. An alternative could be high resolution air ground measures with the implement of low-cost sensors [127,128], however, they are an investment of the local governments, and most of the times is not possible to realize it. An example of where there is insufficient information provided by AQMN stations and a lack of PM10 measures is in Quito, Ecuador [6,129–131], where there is not enough information to establish environmental strategies. Quito, the capital of Ecuador, is a special geographic zone, considering its high elevation altitude (2800 m), in the middle of the Andean region. Considering the difficulties of a city like Quito, one valid alternative to complement AQMN monitoring is applying land use regression models (LUR) [132]. LUR models use different geographical variables as predictors (remote sensing data, meteorological data, road density, vehicular traffic, land use, emission inventory, etc.) [132–135]. However, oftentimes this information cannot be easily accessed. Moreover, these geographical variables are not frequently updated by government institutions. In the case of remote sensing data, the predictors most commonly used in LUR models to retrieve PM10 are aerosol optical depth (AOD) and normalized difference vegetation index (NDVI) from Moderate-Resolution Imaging Spectroradiometer (MODIS) products [136–139]. MODIS products have a low spatial resolution that limits their application in medium or small cities [41,140,141], but they are an efficient alternative to retrieve pollutants in regional (large cities/regions) or national (countries) areas. Consequently, a possible alternative to MODIS products is Landsat data. Nowadays, the operational Landsat satellites are Landsat-7 and Landsat-8 [142,143]. Landsat data have a higher spatial resolution compared with MODIS (30 m instead of 250 m) [141]. Several strategies to retrieve AOD from Landsat data have already been established [142]. Nevertheless, these strategies require AOD ground station data in the study area to have aerosol information in a medium spatial resolution [143,144]. Considering this limiting, other studies suggest that the visible bands of Landsat sensors can be used to invert PM10 [50]. The strategy proposed in this work is useful and effective when the AOD stations are limited.

In order to construct empirical LUR models, some studies have used multiple linear regression (MLR) [144], considering a subset of variables through the stepwise regression (STW) algorithm [26,145]. Nevertheless, the use of MLR cannot analyze the

possible multicollinearity between variables, because we have a high correlation between near bands in the spectrum [146]. Moreover, it is well-known that multicollinearity exists between remote sensing variables [147], producing a source of error in MLR empirical models. However, an alternative which allows the computing of more accurate models, avoiding multicollinearity, is to use partial least square (PLS) regression [34,148,149] or an artificial neuronal network (ANN) [150]. Generally, ANNs give more accurate results in comparison with traditional linear methods, considering the complexity of modeling air pollutants. Some atmospheric studies use a multilayer perceptron (MLP) in the context of ANN in order to obtain a predictor model [144,151].
In Alvarez-Mendoza *et al.* [6], only remote sensing data were considered to compute the LUR model based in a MLR without a method to select predictors. In this work, three main objectives are proposed: (i) using only remote sensing data will be used to establish LUR models without any AOD predictor; (ii) making a comparison between three different remote sensing satellite/sensors (MODIS, Landsat-7 and Landsat-8) to retrieve long-term PM10 considering only a selection of predictors and; (iii) comparing the accuracy of different techniques (STW, PLS and MLP) in the generation of the predictive models. The two last items are the new contributions of this work.

4.3 Materials and Methods

4.3.1. Study Area

The study area is the urban zone of Quito, the capital of Ecuador. Quito comprises 45 urban parishes or *parroquias*, distributed between the latitudes 0º30'S and 0º10'N and the longitudes 78º10'W and 78º40'W (Figure 4.1). The average elevation is around 2800 meters above sea level. The city is located in the middle of the Andean Region. The mean minimum and maximum temperatures are approximately 9.0°C and 25.4°C, respectively. On the other hand, Quito is a region without four seasons because it is in the tropical area, near to the equatorial line. This area was chosen considering the influence of nine AQMN stations.

4.3.2. PM10 data from AQMN stations

In order to monitor air quality in Quito, nine stations have been acquiring air pollutants since 2002 (Figure 4.1). Together they form the "Red Metropolitana de Monitoreo Atmosférico de Quito" (REMMAQ) [38]. REEMAQ is the AQMN of Quito, where one of the air pollutants daily measured is PM10. These data are public and free to download (http://www.quitoambiente.gob.ec/ambiente/index.php/datos-horarios-historicos). The

PM10 concentration is measured in micrograms per cubic meter (μg/m^3). In this study, we use three-month-averages from 2013 to 2017, matching with the dates of the remote sensing data (time when the satellite passes over the study area). The main reasons to use three-month-averages are the few available remote sensing data and REMMAQ stations (stations without data in some months or with negative data values). In this study, PM10 three-month-averages is used as dependent variable.

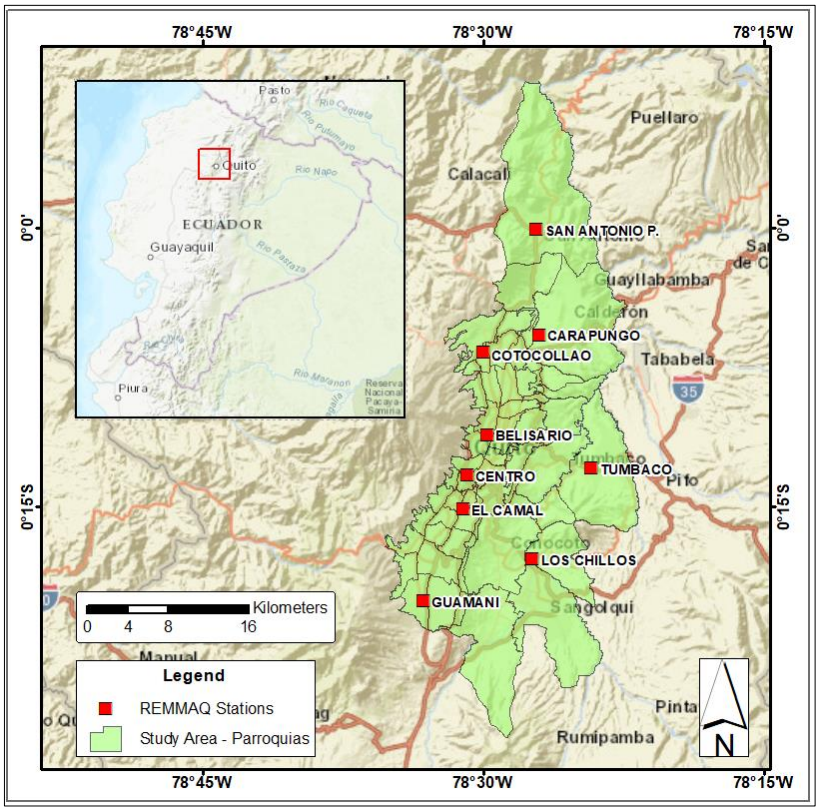


Figure 4.1. Map of the study area (red dots for REEMAQ (Red Metropolitana de Monitoreo Atmosférico de Quito) stations and green polygons for urban parishes).

4.3.3. Remote sensing data predictors

In this study, three different types of remote sensing data were used to retrieve PM10 between 2013 and 2017: Landsat-7 ETM+, Landsat-8 OLI/TIRS and MODIS/Terra and Aqua (Table 4.1). The remote sensing data are free to download from the United States Geological Survey (USGS) website (http://earthexplorer.usgs.gov). Moreover, only images with less than 10% cloud cover were considered in the study, because one of the main problems in these regions is the presence of a high cloud density [32,152]. According to this limitation, just 40% of remote sensing data was considered.

Table 4.1. Characteristics of satellites and sensors used in the study

Satellite	Sensor	Overpass time of satellite	Spatial resolution
Landsat-7	Enhanced Thematic Mapper Plus (ETM+)	16-days	30 meters
Landsat-8	Operational Land Imager (OLI) Thermal Infrared Sensor (TIRS)	16-days	30 meters
Terra (EOS AM-1) Aqua (EOS PM-1)	Moderate Resolution Imaging Spectroradiometer (MODIS) MCD43A4	1 to 2 days	500 meters

The predictors or independent variables (surface reflectance bands and environmental indexes) are listed in Table 4.1. The selection of remote sensing predictors was related to their possible correlation with the PM10 concentration [129,153–155]. In the case of the environmental indexes, the most popular indexes in LUR studies to retrieve PM10 were used. They were computed as (4.1), (4.2), (4.3), (4.4) and (4.5) in Table 4.2, respectively.

Table 4.2. Remote sensing predictors used to build the model for each sensor.

Predictors	Landsat-7	Landsat-8	MODIS
Blue band (B) Green band (G) Red band (R) Near Infrared (NIR) Short Wave infrared (SWIR)	Landsat surface data Level-2	Landsat surface data Level-2	MODIS MOD09A1 MYD09A1 products
Normalized Difference Vegetation Index (NDVI)	$NDVI = \frac{NIR-R}{NIR+R}$ (4.1)		MODIS MOD13Q1 MYD13Q1products
Normalized Difference Soil Index (NDSI)	$NDSI = \frac{SWIR-NIR}{SWIR+NIR}$ (4.2)		
Soil-Adjusted Vegetation Index (SAVI)	$SAVI = (1+L)\frac{NIR-R}{NIR+R+L}$ (4.3) where L represents a minimal change in the soil brightness with a value of 0.5 [5]		

Normalized Difference Water Index (NDWI)	$NDWI = \frac{G-NIR}{G+NIR}$ (4.4)	
Land Surface Temperature (LST)	$LST = \frac{BT}{\left(1+\left(\frac{\lambda * BT}{\rho}\right) ln\varepsilon\right)} - 273.15$ (4.5) where BT is the brightness temperature, λ is the center wavelength (Landsat-7 = 11.45 μm, Landsat-8 = 10.8 μm) [156], ρ is a constant and ε is the emissivity [157,158].	MODIS MOD11A1 MYD11A1 products

4.3.4. LUR models

LUR models are an alternative to predict the spatialization of air pollutants, particularly when the number of AQMN stations is limited. They use different geographical variables such as roads, traffic information, meteorological and remote sensing data and other environmental variables, in order to build a model to retrieve air pollutants. However, often several geographical variables are not available. Thus, we should use simple alternatives, such as free remote sensing data, as variables to approach a LUR model.

In most cases, LUR uses MLR to establish the model [159,160]. MLR allows an easy and simple model construction. In our case, the dependent variable is the quarterly PM10 value and the independent variables or spatial predictors are the remote sensing data in each coordinate of the AQMN station, considering the free cloud pixel value. Equation 4.6 shows the original LUR model, considering all the remote sensing predictors in MLR.

$$PM10 = I + aNDVI - bNDSI - cSAVI + dNDWI - eLST - fB - gG + hR + iNIR + jSWIR + kY - lS \quad (4.6)$$

where I is the intercept, NDVI is Normalized Difference Vegetation Index, NDSI is the Normalized Difference Soil Index, SAVI is the Soil-Adjusted Vegetation Index, NDWI is the Normalized Difference Water Index, LST is the Land Surface Temperature, B is the blue band, G is the green band, R is the red band, NIR is the near infrared band, SWIR is the shortwave infrared band, Y is the year of image acquisition, S is the three-month-averages of image acquisition (January–March - 1, April–June - 2, July–September - 3, and October–November - 4), a, b, …, l, are the coefficients in each predictor. The other variables are described in Table 4.2.

Nevertheless, considering that multicollinearity exists between remote sensing variables [147], different predictor techniques should be employed to compute the LUR model. We compare three techniques, namely, MLR with STW, PLS and ANN, in order to find the fittest model (Figure 4.2).

In the first model, we use MLR considering an STW. It contemplates different parameters in order to identify the most adequate/influencing variables as predictors. The parameters used to subset the variables are: (i) the residual sum of squares for each model (RSS); (ii) the adjusted regression coefficient R^2 (Adj. R2); (iii) Mallows' Cp (CP) and; (iv) Bayesian information criterion (BIC).

The second model uses PLS with the STW criteria to select the predictors. The main challenge when using PLS is to avoid multicollinearity, finding an alternative when we have few data and a significant number of predictors [76]. PLS generates new latent variables or components in a lineal way.

Finally, the last model uses an ANN in an MLP, with a hidden layer and six hidden nodes to compute the predictive model. The nodes are computed according to [161]. In this model we use all the predictors. This method is used when the model is complex, giving a different weight to each predictor corresponding to its importance. Additionally, we use a non-linear activation function with backpropagation. The training data to build the MLP consider 75% of the dataset and the rest 25% for test. We use a backpropagation approach to train the algorithm. The R studio software was used in this study to extract the data and to compute all the models.

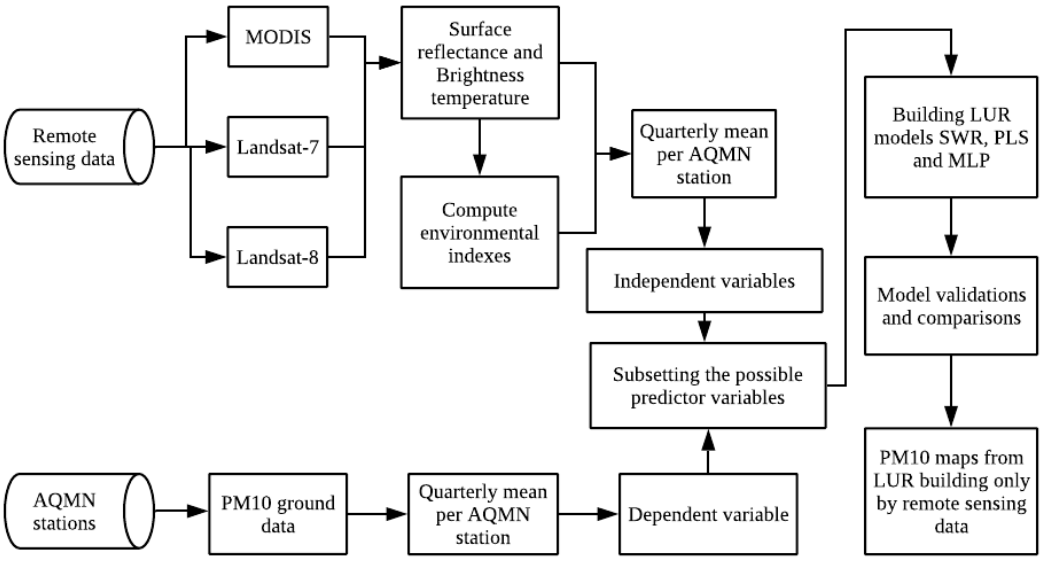


Figure 4.2. Workflow of the methodology proposed to establish the land use regression (LUR) models.

4.4 Results

PM10 ground measurements and remote sensing data are matched in a table with the same date. Thus, the unique condition is to consider remote sensing data with less than 10% cloud density. So, the three-month-averages matching tables for each sensor

contain 35 observations for Landsat-7, 93 observations for Landsat-8 and 108 observations for MODIS. The main reasons to have only these numbers of observations are the high cloud density in the study area and the incomplete/not available air pollution data. Furthermore, the criteria to select predict variables consider 5 dependent variables for Landsat-7, 8 dependent variables for Landast-8 and 6 dependent variables for MODIS, for each STW and PLS model, as shown in table 4.3. They were obtained according to STW criteria (RSS, Adj. R2, CP and BIC). The variables common to all the three cases considered are blue band, near infrared (NIR) band and Normalized Difference Vegetation Index (NDVI).

Table 4.3. Number of observations and predictors per satellite to build the LUR models.

Variable	Landsat-7	Landsat-8	MODIS
No. Observations	35	93	108
No. Predictors	5	8	6
Predictors	NDVI B R NIR S	NDVI SAVI LST B G R NIR Y	NDVI B G R NIR S

The LUR models are computed considering STW and PLS regressions in a linear way and MLP in a non-linear way. They are shown and compared in Table 4.4 (Equations 4.7 to 4.12). In the case of Landsat-7, the STW shows a coefficient of determination (R^2) of 0.37, the PLS a R^2 of 0.36, and, for MLP, a R^2 of 0.46. The lowest root-mean-square error (RMSE) was obtained for STW with a value of 9.47. For Landsat-8, in STW a R^2 of 0.42 was obtained, and a R^2 of 0.43 for PLS, and a R^2 of 0.68 for MLP (Figure 4.3). The lowest RMSE obtained was for MLP. Finally, for MODIS, a R^2 of 0.15 for STW, a R^2 of 0.19 for PLS and a R^2 of 0.25 for MLP were obtained. The lowest RMSE was for STW.

Table 4.4. LUR models for each sensor with different regression techniques. In the case of MLP, the model is not linear.

Sensor	Model	Equation/Method	Coefficient of determination (R^2)	Root-mean-square error (RMSE)
Landsat-7 ETM+	Stepwise regression (STW)	$\mathrm{PM10} = -26.770 + 205.289\mathrm{NDVI} - 0.073\mathrm{B} + 0.144\mathrm{R} - 0.048\mathrm{NIR} + 2.270\mathrm{S}$ (4.7)	0.37	9.47
	Partial least square regression (PLS)	$\mathrm{PM10} = 24.786 - 54.369\mathrm{NDVI} - 0.059\mathrm{B} + 0.049\mathrm{R} - 0.008\mathrm{NIR} + 2.165\mathrm{S}$ (4.8)	0.36	10.14
	Multilayer perceptron (MLP)	Non-linear. One hidden layer and six hidden nodes.	0.46	12.69

Landsat-8 OLI/TIRS	STW	$PM10 = -4125.506 + 350.130NDVI - 200.334SAVI — 0.936LST - 0.035B - 0.036G + 0.099R - 0.013NIR + 2.061Y$ (4.9)	0.42	9.19
	PLS	$PM10 = -4146.508 + 115.816NDVI - 40.465SAVI — 1.020LST - 0.036B - 0.038G + 0.104R - 0.016NIR + 2.073Y$ (4.10)	0.43	9.46
	MLP	Non-linear. One hidden layer and six hidden nodes.	0.68	6.22
MODIS	STW	$PM10 = 1.248 + 93.411NDVI + 0.056B - 0.070G + 0.056R - 0.017NIR + 3.190S$ (4.11)	0.15	12.91
	PLS	$PM10 = 5.661 + 79.106NDVI + 0.060B - 0.072G + 0.050R - 0.014NIR + 3.308S$ (4.12)	0.19	12.93
	MLP	Non-linear. One hidden layer and six hidden nodes.	0.25	16.38

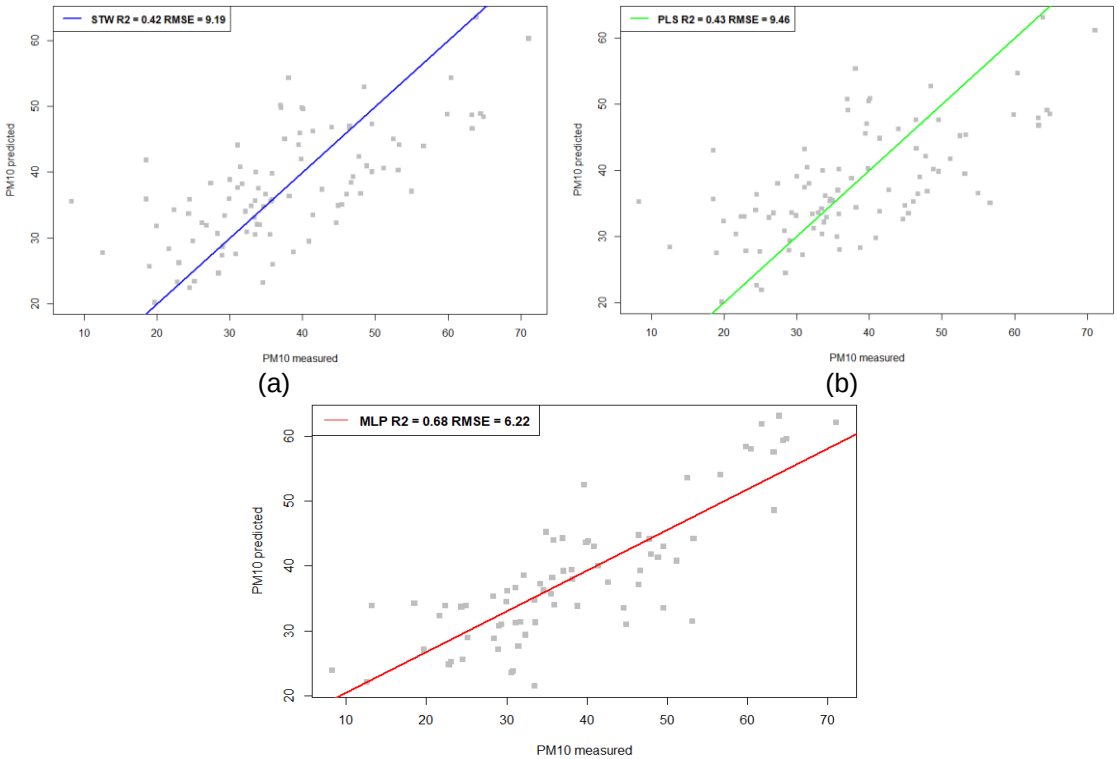


Figure 4.3. Comparison between R^2 and RMSE in the model results for Landsat-8 data: (a) STW; (b) PLS; (c) MLP.

The results in Table 4.3 show that Landsat-8 data with MLP are the fittest model. The MLP employed (Figure 4.4) has one hidden layer with six hidden nodes.

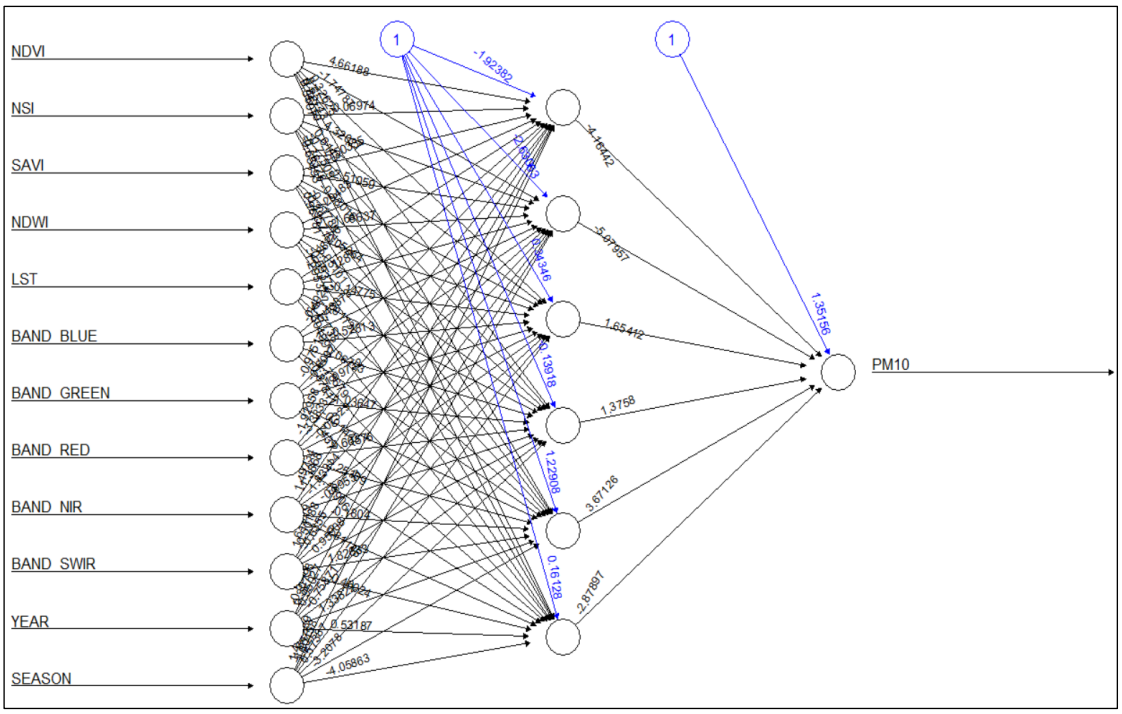


Figure 4.4. MLP diagram for Landsat-8 data.

Figure 4.5 shows the relative variable importance according to the assigned weights, where the red band is the most significant in the model, while LST presented the lowest significance.

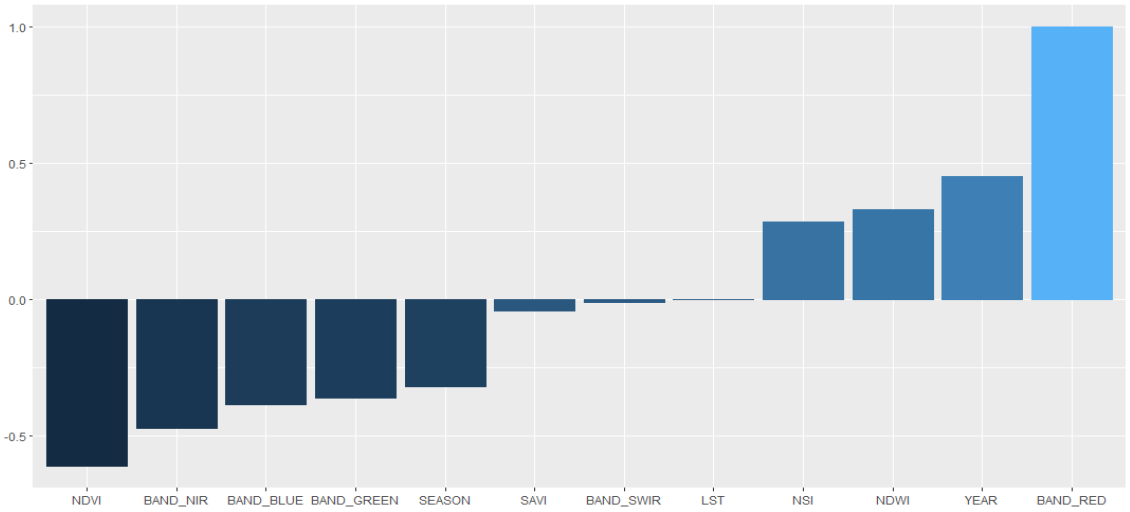
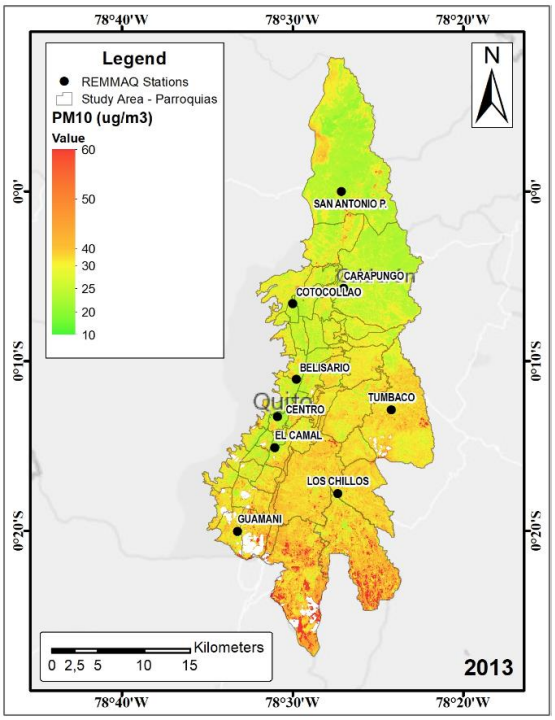
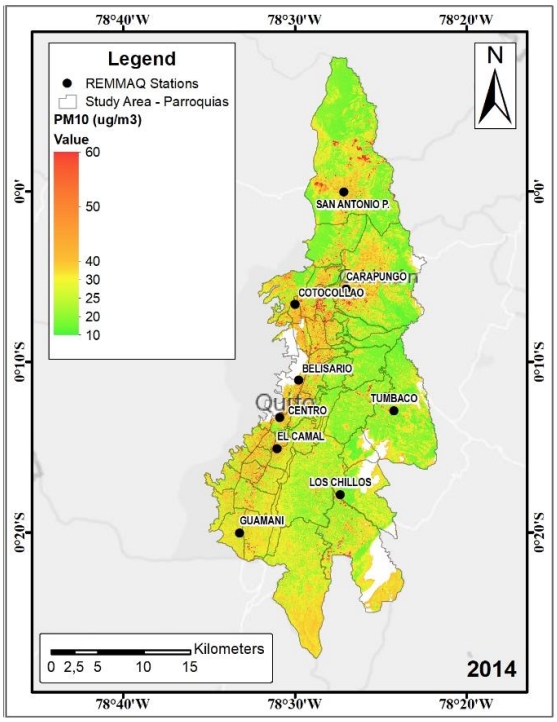


Figure 4.5. Relative variable importance in Landsat-8 MLP. The scale is between -0.5 and 1, where 0 is the lowest (null) importance.

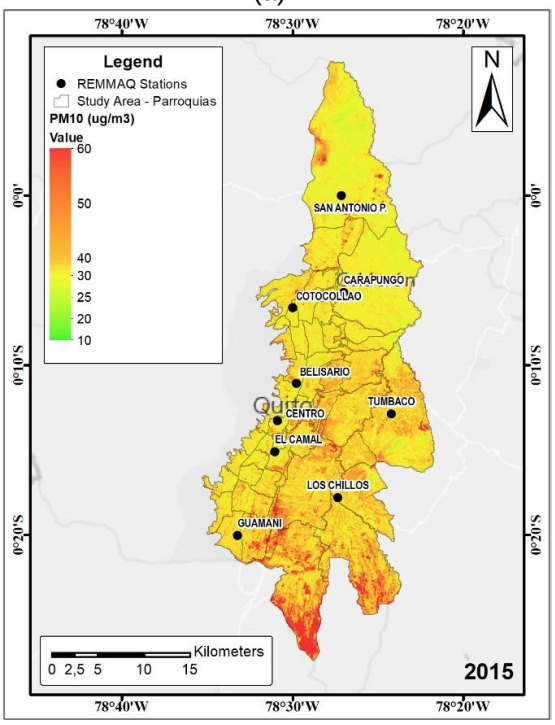
The Landsat-8 LUR-MLP model is chosen to predict PM10, considering the highest R^2 and the lowest RMSE. In Figure 4.6, the quarterly maps show the PM10 spatial concentration during 2015, in a color scale in $\mu g/m^3$. The white gaps showed in the maps are clouds with a high density.


Legend
REMMAQ Stations
Study Area - Parroquias
PM10 (ug/m3)
Value
60
50
40
30
25
20
10
N
SAN ANTONIO P.
CARAPUNGO
COTOCOLLAO
BELISARIO
TUMBACO
CENTRO
EL CAMAL
LOS CHILLOS
GUAMANI
Quito
Kilometers
0 2,5 5 10 15
2013
78°40'W
78°30'W
78°20'W
0°0'
0°10'S
0°20'S

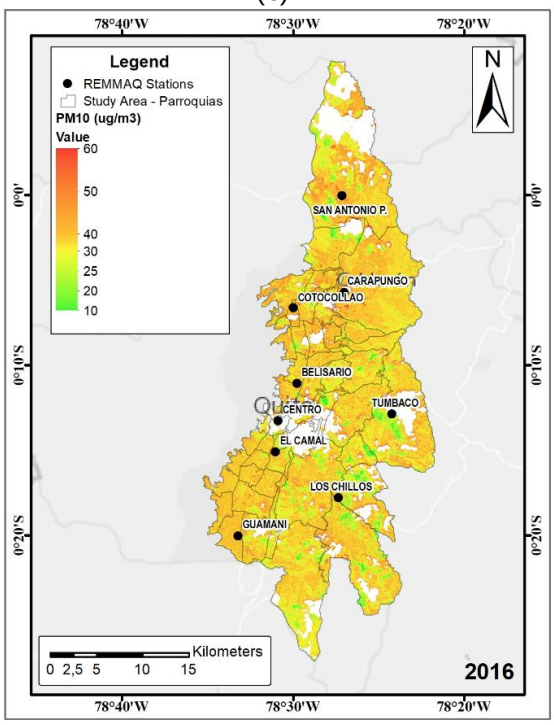
(a)


Legend
REMMAQ Stations
Study Area - Parroquias
PM10 (ug/m3)
Value
60
50
40
30
25
20
10
N
SAN ANTONIO P.
CARAPUNGO
COTOCOLLAO
BELISARIO
TUMBACO
CENTRO
EL CAMAL
LOS CHILLOS
GUAMANI
Quito
Kilometers
0 2,5 5 10 15
2014
78°40'W
78°30'W
78°20'W
0°0'
0°10'S
0°20'S

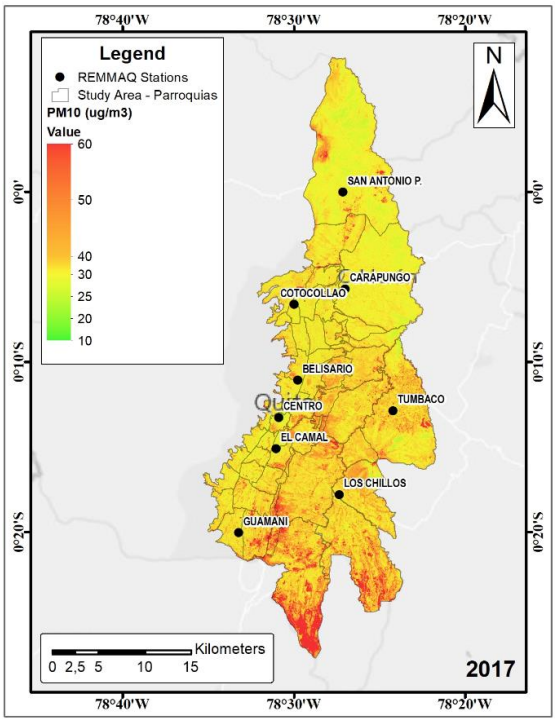
(b)


Legend
REMMAQ Stations
Study Area - Parroquias
PM10 (ug/m3)
Value
60
50
40
30
25
20
10
N
SAN ANTONIO P.
CARAPUNGO
COTOCOLLAO
BELISARIO
TUMBACO
CENTRO
EL CAMAL
LOS CHILLOS
GUAMANI
Quito
Kilometers
0 2,5 5 10 15
2015
78°40'W
78°30'W
78°20'W
0°0'
0°10'S
0°20'S

(c)


Legend
REMMAQ Stations
Study Area - Parroquias
PM10 (ug/m3)
Value
60
50
40
30
25
20
10
N
SAN ANTONIO P.
CARAPUNGO
COTOCOLLAO
BELISARIO
TUMBACO
CENTRO
EL CAMAL
LOS CHILLOS
GUAMANI
Quito
Kilometers
0 2,5 5 10 15
2016
78°40'W
78°30'W
78°20'W
0°0'
0°10'S
0°20'S

(d)



(e)

Figure 4.6. PM10 concentrations during the season 4 (July to September) with Landsat-8 LUR-MLP model in: (a) 2013; (b) 2014; (c) 2015; (d) 2016; (e) 2017. The white gaps represent areas with a high cloud density.

4.5 Discussion

As demonstrated in this study, LUR models are an interesting alternative to model air quality, specifically PM10 concentrations, when the in-situ air quality measures are insufficient. Usually, most of the predictors are geographical variables (such as roads), traffic, meteorological data, and others [132]. LUR models are usually applied in small cities or regions where AQMN stations are limited [162], and where spatial interpolation techniques, such as ordinary kriging or inverse distance weighting, cannot be applied, considering the low number of ground measurements available [163]. One of the main problems with these geographic variables is the low accessibility to the data and the time of acquisition. Sometimes, these variables are obsolete, and they are not enough to establish a possible trend.

In this study, we propose an alternative, considering only free remote sensing variables. We apply this approach to the city of Quito, Ecuador, during the period between 2013 and 2017, in order to compare three different satellite data. Quito is growing in new poles. When REEMAQ was established in 2002, Quito did not have its current size and configuration. Now, REEMAQ is an obsolete air quality network, especially in the distribution of stations, which urgently needs improvement. Air pollutant spatial models are techniques based on interpolation or geostatistics approaches, which can be useful if a reasonable number of stations are available with a good spatial distribution [164]. In

this study, only nine stations are available. Moreover, in some cases the data are incomplete during some months. Additionally, according to some authors [127,128], it is possible to have more air ground data with low-cost sensors, however they must be implemented in the cities in order to monitor the air quality. The alternative to improve the air quality model in Quito is to establish different spatiotemporal LUR models, considering only remote sensing data as predictor variables. A preliminary study shows the use of only remote sensing variables, but using a MLR in order to build the model. The limitation is the use of all remote sensing predictors without considering the collinearity [6]. In order to establish the models, three different remote sensing data were tested (Landsat-7, Landsat-8 and MODIS) and three techniques for modeling (STW, PLS and MLP) were employed. The selected variables to compute the model are the visible NIR and SWIR bands of the three sensors, different environmental indices (NDVI, NDSI, SAVI, NDWI) and LST, computed from the data retrieved from each sensor. Most of the studies published use aerosol optical thickness (AOT) derived from MODIS (MOD04) [165] as the input in LUR models, however, this product has a low spatial resolution (3 x 3 km)[166]. This resolution is not practicable when considering cities like Quito, where the maximum width is near to 10 km. On the other hand, some MODIS products do not have a suitable quality for local studies [167]. Other studies use Landsat-8 combined with AOT ground stations to spatially model the AOT [142]. This could be a good alternative, however, in our study area we do not have access to this information between 2013 and 2017.

Comparing the LUR models established, we found that Landsat-8 is the most adequate sensor to model PM10 concentration, considering the 93 records and according to a previous study [6]. MLP is the fittest alternative to model PM10, with a R^2 of 0.68 and a RMSE of 6.22. In this context, the non-linear model (MLP) has a fitter result when compared to the linear models (STW and PLS) [144]. Therefore, the LUR-MLP model was chosen to map the spatial concentration of PM10 in Quito, between 2013 to 2017. MODIS presents the lowest R^2 with a value of 0.19, considering the PLS regression. This could be related to the lowest spatial resolution. Thus, most of the LUR models use MLR or STW. MLR is easy to implement. However, one of the main problems could be the multicollinearity, because MLR not analyze the correlation between predictors [168]. On the other hand, the linear PLS helps to avoid the multicollinearity creating new latent variables with few observations [34]. In a future work, a possible combination between STW (in order to select the predictor variables), non-linear PLS (in order to avoid the multicollinearity between remote sensing data) and a machine learning technique (as ANN) can improve the LUR models [169].

In the case of the predictors, all the models present, in all the cases, the variables blue band, NIR and NDVI. In the case of NDVI, a possible reason is the direct influence of vegetation on the PM10 concentration and distribution [137]. On the other hand, the red band has the most importance in MLP, because there could be a relationship between the retrieval of PM10 with the blue and red bands [50]. In most of the LUR studies, the authors use traffic, roads, meteorological, land use, population and other predictors, reporting values of R^2 according to the reality of each local [144]. These models also considered different time periods (monthly, quartly, yearly). The main difference of our approach is the use of remote sensing data only as predictors, which can replace the necessity to have all geographical variables. Another advantage is the data availability and continuity in order to recompute the LUR models. One of the main limitations of our model is the high cloud density presented in the images during all the year [32], making complicated to use more data in order to improve the model. However, in a future work will intend to have more satellite sensors or to find new alternatives to recover remote sensing data contaminated with clouds [33].

Figure 6 shows variations year by year according to PM10 mean concentration based on in-situ data (REEMAQ Stations). We choose the 3th season to show the variation year by year (2013 - 2017), because we have more remote sensing data available (without a high cloud density) during this time-window. According to the results presented in Figure 6, an increasing of PM10 concentration between 2013 to 2017 is notorious in the most of the urban parishes [170]. However, some areas showed a decreasing tendency in some years. The lowest PM10 concentration was found in some peripheral parishes during the 2014 year, because the air stations which influences these parishes (Tumbaco and Los Chillos) had a variation in the concentrations. Thus, Tumbaco and Los Chillos stations are in the east part of the study area and began to present the lower values in 2014 followed by 2013, according to the in-situ measures. After 2014, the PM10 values for these stations began to increase. The main reason could be related to the begin operation of the new airport of Quito (2013), and the construction of important road infrastructures around it (end of 2014). Another possible explication is the traffic influence during the last years, particularly in the peripheral areas were an increment was registered since 2015 and also the increase of the population in these areas [171]. In the northern parishes, the stations of San Antonio P. and Carapungo are influenced by the presence of stone and sandy point quarries [172]. The stations Centro, Belisario and El Camal are in the city downtown, and it is the main reason why an increase of PM10 concentration during the last years is verified in the centre parishes.

According to our results, several areas presented concentrations higher than 50 µg/m^3 (Figure 4.6), while the World Health Organization (WHO) recommends, in its guidelines, maximum values of 20 µg/m^3 as an annual mean and 50 µg/m^3 as a 24-hour mean [123]. However, some areas do not show values, due to the high cloud density (white areas in Figure 4.6). Thus, the PM10 concentration maps from the Landsat-8 LUR-MLP model can help local government decision makers to manage air quality concentration and to organize new policies, specifically in the places where the highest concentrations were identified.

4.6 Conclusions

In this study, three different satellite datasets were compared to retrieve models of PM10 through LUR, in Quito, Ecuador between 2013 and 2017. Additionally, three techniques were compared to compute the LUR models (SWR, PLS and MLP). From this work, several conclusions could be taken: (i) it is possible to build empirical models established only using remote sensing variables as predictors without any other geographic variables, as traditional LUR models; (ii) in the case of Quito, the study results show that Landsat-8 provides the most suitable satellite data to retrieve PM10, in comparison with Landsat-7 and MODIS; (iii) MLP allows the obtainment of the most robust result in comparison with the other modeling techniques. MLP is the fittest alternative to model PM10, with a R^2 of 0.68 and a RMSE of 6.22, and; (iv) the MLP model established helps in the spatial mapping of PM10, where in the time window of this study, were found areas with PM10 values higher than the limit established by WHO. Thus, these models are useful in the management of air quality in the city of Quito and can be applied to other locations with similar characteristics.

5. Article 3: Spatial estimation of surface ozone concentrations in Quito Ecuador with remote sensing data, air pollution measurements and meteorological variables

Cesar I. Alvarez-Mendoza[1,2*], Ana Teodoro[1,3], and Lenin Ramirez-Cando[2]

[1] University of Porto, Department of Geosciencies, Environment and Land Planning, Faculty of Sciences, Rua Campo Alegre 687, Porto 4169-007, Portugal; calvarezm@ups.edu.ec

[2] Universidad Politécnica Salesiana, Grupo de Investigación Ambiental en el Desarrollo Sustentable, Carrera de Ingeniería Ambiental, Quito, Ecuador; lramirez@ups.edu.ec

[3] Earth Sciences Institute (ICT), Pole of the FCUP, University of Porto, Porto, Portugal; amteodor@fc.up.pt

Academic Editor: G. B. Wiersma

*Received: 5 September 2018 / Accepted: 30 January 2019 / Published: 11 February 2019*Journal: **Environmental Monitoring and Assessment. Elsevier.**

March 2019, 191:155; doi.org/10.1007/s10661-019-7286-6

5.1 Abstract

Surface ozone is problematic to air pollution. It influences respiratory health. The air quality monitoring stations measure pollutants as surface ozone, but they are sometimes insufficient or do not have an adequate distribution for understanding the spatial distribution of pollutants in an urban area. In recent years, some projects have found a connection between remote sensing, air quality and health data. In this study, we apply an empirical land use regression (LUR) model to retrieve surface ozone in Quito. The model considers remote sensing data, air pollution measurements and meteorological variables. The objective is to use all available Landsat-8 images from 2014 and the air quality monitoring station data during the same dates of image acquisition. Nineteen input variables were considered, selecting by a stepwise regression and modelling with a partial least square (PLS) regression to avoid multicollinearity. The final surface ozone model includes ten independent variables and presents a coefficient of determination (R^2) of 0.768. The model proposed help to understand the spatial concentration of surface ozone in Quito with a better spatial resolution.

Keywords: Landsat-8, Quito, Ozone, PLS, Air modelling

5.2 Introduction

Surface ozone (O_3) is one of the principal greenhouse gases [173]. It is produced in the troposphere and is not emitted directly into the air. A chemical reaction between nitrogen oxides (NO_x), volatile organic compounds (VOC) and sunlight produces O_3 [174]. Thus, urban growth, vehicular traffic and industry are sources of NOx and VOC in cities, deteriorating the vegetation conditions [175], the air quality and creating a health problem [15,176].

Several cities around the world have an air quality monitoring network (AQMN) to manage air pollution [62,177]. One of the cities with an AQMN is Quito, the capital of Ecuador. The city has traffic and population problems that increase air pollution. Its AQMN is the “Red Metropolitana de Monitoreo Atmosférico de Quito” (REMMAQ), constituted by nine stations. It has managed the air quality in Quito in real time since 2002 [38]. The REMMAQ stations measure air pollutants such as carbon monoxide (CO), nitrogen dioxide (NO_2) as part of NO_x, sulphur dioxide (SO_2), particulate matter less than 10 microns (PM10), fine particles less than 2.5 microns (PM2.5) and O_3. Nevertheless, the number of stations is insufficient to measure the air quality in all urban zones in the city.

Some empirical models to retrieve the spatial concentration of air pollutants have been developed using variables such as roads information and vegetation. The land use regression (LUR) models are the basis of most of these approaches. The principle of LUR focuses on the environmental characteristics of the place where the pollutant is present [159]. Some models consider remote sensing data, meteorological data (MD), aerosol optical depth (AOD) field measurements and AQMN data [19,178,179]. In most of these studies, the limitations are related to the input variables, especially AOD field measurements. This is because models require AOD parameters to obtain high-resolution spatialization [142,179]. The most commonly used remote sensing data are Landsat [20,180,181] and MODIS [133,182] sensors. The main advantage of Landsat images in specific Landsat-8 [10], is the high spatial resolution to map middle cities. Their limitation is the temporal resolution (16 days) [10]. The advantage of MODIS is its high temporal resolution, but the major limitation is the low spatial resolution, which limits the accurate retrieval of maps (Daac, Falls, & March 2012). Moreover, remote sensing data are used to obtain environmental variables such as vegetation health [184,185] to input variables in the air pollutant models. Furthermore, empirical models using remote sensing data are focused on only some air pollutants, such as NO_2, PM10 and PM2.5. At present, the main challenge is to retrieve the remaining air pollutants, such as O_3, which is considered only in few studies [186].

In the case of Quito, a study found the spatial distribution of PM10 by applying remote sensing data [129]. The main limitation of the study was the small quantity of data used (3 images). On the other hand, a study making a comparison between remote sensing to retrieve air pollutant in Quito is considered [6]. However, there are few studies about air quality in the city, specifically considering O_3 [187]. Thus, the possibility of obtaining AQMN public data, and combining them with other environmental variables, can lead to new models for retrieving air pollutants in places where AQMN are insufficient.

This study uses remote sensing data, air pollution measurements and meteorological variables to retrieve O_3 for one year (2014) in Quito. Moreover, this study combines two regression techniques, stepwise regression (SWR) and partial least-square regression (PLS), to compute the O_3 model, finding the fittest model to spatialize the variable in all the areas. The main objective is to find the spatial variables that influence O_3 in Quito.

5.3 Materials and Methods

5.3.1. Study Area

This study was developed in Quito, the capital of Ecuador. The city elevation is approximately 2800m over sea level. During 2014, the mean minimum and maximum temperatures were 9.0°C and 25.4°C [111]. Furthermore, Quito has a dry season and a wet season. It does not have four seasons considering that the city is in the middle of the tropic zone. The latitude and longitude of the study area are 0°30'S to 0°10'N and 78°10'W to 78°40'W. These coordinates delimit most of the urban zone, which is divided into urban parishes (Figure 5.1).

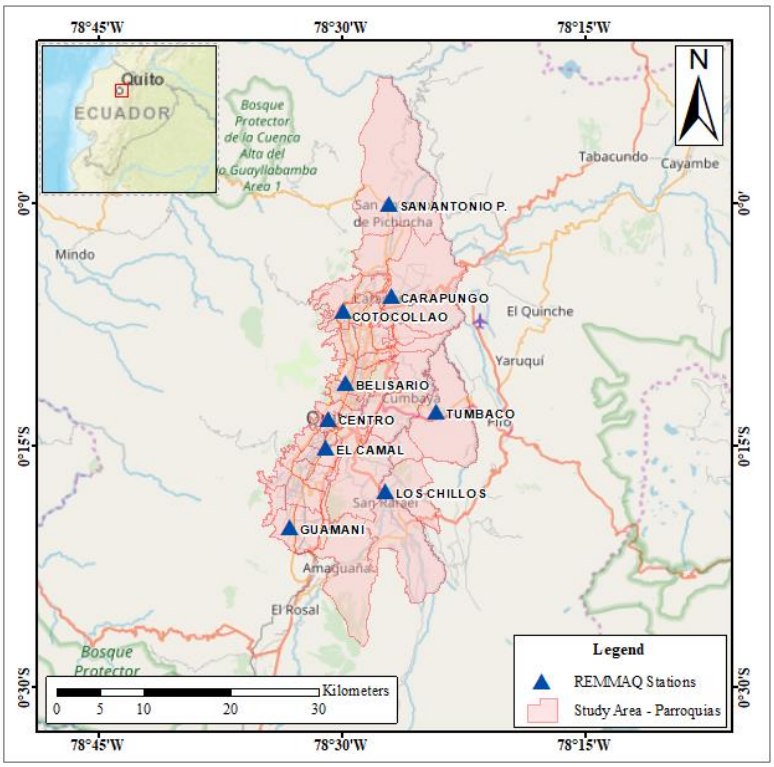


Figure 5.1. Quito's urban parishes considered as the study area. The blue marks represent the REMMAQ stations.

5.3.2. Air pollutant ground data

The daily air pollutant concentration data from 2014 were obtained from the REMMAQ stations. The REMMAQ has nine automatic stations that have been operated by the "Secretaria del Ambiente de Quito" since 2002 (Figure 5.1). The stations measure concentrations of air pollutants such as PM2.5, SO_2, CO, O_3, NO_2, PM10 and MD (Table 5.1). In this study, daily average measurements were considered to match with the satellite overpass (Figure 5.2) (*See section 5.3.4*). Furthermore, only complete datasets were used, which means that if a dataset was incomplete, it was not considered for the model establishment. PM2.5, SO_2, CO, and NO_2 were the complete datasets to estimate O_3. The pollutant concentration was measured in micrograms per cubic meter ($\mu g/m^3$) according to the Environmental Protection Agency (EPA) methods. The O_3 measuring device was a Teledyne API/T400, and the collection method was EPA No. EQOA-0992-

087 [38]. The hourly pollutant concentration data have public access (http://www.quitoambiente.gob.ec/ambiente/index.php/datos-horarios-historicos).

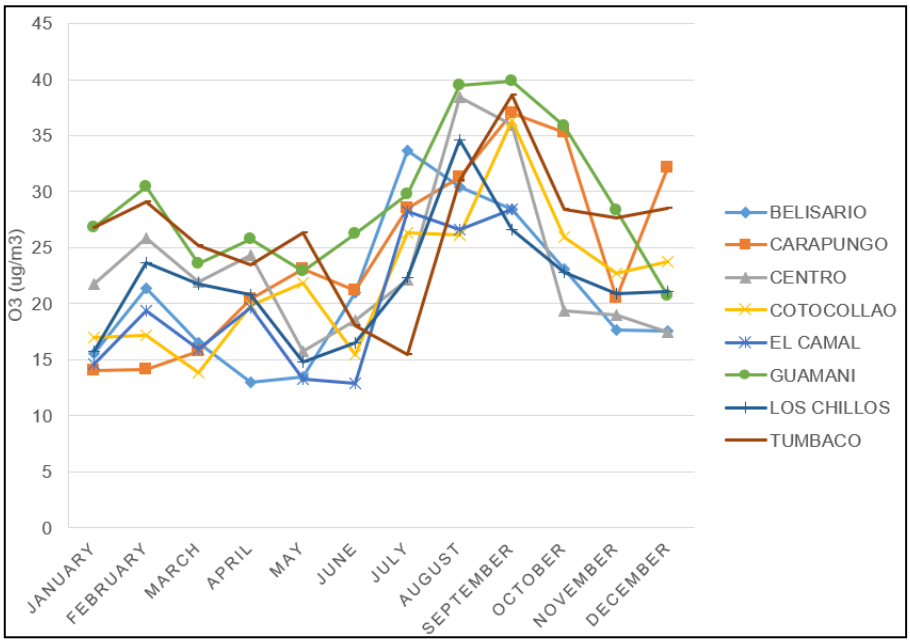


Figure 5.2. Mean levels from 10:00 to 11:00 (GMT-5) of O_3 concentration (µg/m^3) observed in each month during 2014. The San Antonio P. station did not present measures during 2014.

5.3.3. Meteorological data

The MD were collected only by eight REMMAQ stations (Table 5.1). The data used were the daily average temperature (TMP) in Celsius degrees (°C), relative humidity (HM) in percentage (%) and solar radiation (SR) in Watt per square metres (W/m^2). The precipitation measurements were not used because most of the values were null in the time range considered.

In both cases, (air pollutant ground data and meteorological data), the R software was used to analyse the data and compute the statistics. The packages *readxl* and *stringi* were used.

Table 5.1. Field sensors of the REEMAQ

Station	Variables measured
Cotocollao	PM2.5, SO_2, CO, O_3, NO_2, PM10, MD
Carcelen	PM2.5, SO_2, CO, O_3, NO_2, PM10, MD
Belisario	PM2.5, SO_2, CO, O_3, NO_2, MD
Jipijapa	PM2.5, SO_2, CO, O_3, NO_2, PM10, MD
Camal	PM2.5, SO_2, CO, O_3, NO_2, MD
Centro	PM2.5, SO_2, CO, O_3, NO_2
Guamani	SO_2, CO, O_3, NO_2, PM10, MD
Tumbaco	SO_2, O_3, PM10, MD
Los Chillos	PM2.5, SO_2, CO, O_3, NO_2, MD

5.3.4. Remote sensing data

Landsat-8 is a satellite launched on February 11, 2013. It is the last satellite of the Landsat project launched. The satellite carries two push-brown instruments to collect land remote sensing data on an image: The Operational Land Imager (OLI) with 9 bands and the thermal infrared sensors (TIRS) with two bands. Additionally, the Landsat-8 data file provides a quality assessment band (QA) to assess the different image products. The Landsat-8 images are freely available on the United States Geological Survey (USGS) website. The USGS develops research-quality and application-ready products such as the Landsat-8 surface reflectance Level-2 products (L2T). These products are generated from the Landsat Surface Reflectance Code (LaSRC) [50]. The LaSRC products are radiometric and atmospherically corrected. The LaSRC products include surface reflectance of the OLI bands (bands 1 to 9), top-of-atmosphere brightness temperature (BT) (band 10 and band 11) and some environmental indexes such as the normalized difference vegetation index (NDVI), soil-adjusted vegetation index (SAVI) and enhanced vegetation index (EVI).

In this study, Landsat-8 L2T images were downloaded from the Earth Resources Observation and Science (EROS) Center Science Processing Architecture (ESPA) at the demand interface (https://espa.cr.usgs.gov/). The search criteria were images in 2014 with less than 20% cloud cover in the study area. One of the challenges was to choose the subset of images without high cloud density in the study area [152]. According to the search criteria, ten images (path 11; row 60) were selected (Table 5.2).

Table 5.2. Landsat-8 L2T images selected

No.	Image	Date
1	LC08_L1TP_010060_20140115_20170426_01_T1	15/01/2014
2	LC08_L1TP_010060_20140131_20170426_01_T1	31/01/2014
3	LC08_L1TP_010060_20140216_20170425_01_T1	16/02/2014
4	LC08_L1TP_010060_20140304_20170425_01_T1	04/03/2014
5	LC08_L1TP_010060_20140405_20170424_01_T1	05/04/2014
6	LC08_L1TP_010060_20140608_20170422_01_T1	08/06/2014
7	LC08_L1TP_010060_20140710_20170421_01_T1	10/07/2014
8	LC08_L1TP_010060_20140726_20170420_01_T1	26/07/2014
9	LC08_L1TP_010060_20140811_20170420_01_T1	11/08/2014
10	LC08_L1TP_010060_20141030_20170418_01_T1	30/10/2014

Considering the direct influence of the sunlight over O_3 concentration [188] and knowing the principle of passive remote sensing data to capture the radiation measured reflectance sunlight [88,189], bands 1 to 7 (visible and infrared bands) [10] were used as input variables. NDVI, SAVI and EVI were used to highlight the vegetation because there is a high relation between O_3 and vegetation [190]. The indexes were obtained from

LaSRC and multiplied by 0.0001 [191] to retrieve the surface environmental indexes (values between -1 and 1).

The NVDI provides information about health vegetation, using band 4 (B4) and band 5 (B5) in Landsat-8 images. It is computed using Equation 5.1;

$$NDVI = \frac{B5 - B4}{B5 + B4} \tag{5.1}$$

The SAVI is an improvement of NDVI considering a soil correction factor (usually LS=0.5). Considering Landsat 8, it uses B4 and B5 as input (Equation 5.2).

$$SAVI = (1 + LS)\frac{B5 - B4}{B5 + B4 + LS} \tag{5.2}$$

The EVI enhances the vegetation in areas with high biomass. Thus, EVI helps to identify stress vegetation using Equation 5.3.

$$EVI = G * \frac{B5 - B4}{B5 + C1 * B4 - C2 * B2 + L} \tag{5.3}$$

where the gain factor (G) is 2.5, L is the canopy background adjustment (L=1), C1 and C2 are coefficients for atmospheric resistance (C1=6, C2=7.5). The B4 and B5 have a high contrast in the detection of built-up areas and bare lands areas [57].

Moreover, the land surface temperature (LST) retrieved from remote sensing has been used in other studies to estimate the air quality [20]. It was computed as a function of BT. Equation 5.4 represents the LST in degrees Celsius.

$$LST = \frac{BT}{\left(1 + \left(\frac{\lambda * BT}{p}\right) ln\mathrm{E}\right)} - 273.15 \tag{5.4}$$

where λ is the centre wavelength (λ=10.8 μm), p is a constant obtained in Equation 5.5, E is the emissivity as Equation 5.6 and 273.15 is the value to transform degrees Kelvin to degrees Celsius.

The constant p is estimated using Equation 5.5, where h is the Planck constant (6.626e-34 Js), c is the speed of light (2.998e8 m/s), and s is the Boltzmann constant (1.38e-23 J/K).

$$p = \frac{h * c}{s} \tag{5.5}$$

Equation 5.6 represents the emissivity E [59]. E is the efficiency that a surface emits heat as thermal infrared (TIR) radiation [60].

$$E = \begin{cases} E_s, & NDVI < NDVI_s \\ E_s + (E_v - E_s)P_V, & NDVI_s \leq NDVI \leq NDVI_v \\ E_v, & NDVI > NDVI_v \end{cases} \tag{5.6}$$

where E_s represents the emissivity for soil. A value of 0.973 is used in this study [157]. E_V is the vegetation emissivity with a value of 0.985 in this study [157]. $NDVI_v$ is the NDVI in vegetation with a value of 0.2 [59], $NDVI_s$ is the NDVI in the soil with a value of 0.5 [59] and P_V is the proportion of vegetation in the area using Equation 5.7.

$$P_V = \left(\frac{NDVI - NDVI_s}{NDVI_v - NDVI_s}\right)^2 \quad (5.7)$$

The remote sensing variables were represented as raster data (GeoTIFF format). They were computed in R studio software with the *rgdal* and *raster* packages. Through the shapefile of REMMAQ stations, the raster values for each station were extracted. The package *dismo* was used to perform this task.

5.3.5. Model building

The first step in building the model is the compilation of all possible variables (air measurement data, meteorological data and remote sensing data) in a database. Each row in the table has all the values of these variables in a REMMAQ station during the date established (Table 5.3).

Table 5.3 Variables considered in the model

No.	Variable	Units
Air pollutants ground data	O_3, PM2.5, SO_2, CO, NO_2	$\mu g/m^3$
Meteorological data	Temperature (TMP)	°C
	Relative humidity (HUM)	%
	Solar radiation (SR)	W/m^2
Remote sensing data	Band 1 (B1), Band 2 (B2), Band 3 (B3), Band 4 (B4), Band 5 (B5), Band 6 (B6), Band 7 (B7)	Surface reflectance
	Environmental Indexes: NDVI, SAVI, EVI	-
	Land surface temperature (LST)	°C

LUR models are a good alternative for finding the spatial location of pollutants [192]. LUR are empirical regression models that consider the pollutant of interest as the dependent variable and other geographical variables as independent variables (meteorological data, traffic, topography, remote sensing data, etc.). In this study, we generate an LUR model using the available data from each station on different dates during 2014 to preserve the accuracy of the variables.

Assuming that multicollinearity between variables is real, especially between remote sensing variables [147], a preliminary correlation analysis was realized to provide an overview of which variables are more adequate for integration into the model.

To select the fittest predictor variables and the best model to predict O_3, a subset analysis is performed with stepwise regression. The subset analysis used four analyses: the residual sum of squares for each model (RSS), the adjusted regression coefficient R^2 (Adj. R^2), Mallows' Cp (CP) and the Bayesian information criterion (BIC). The R-package used to compute this was *leaps*.

The original LUR model with all the possible predictor variables as input in the analysis is shown in Equation 5.8.

$$O_3 = aPM2.5 + bSO_2 + cCO + dNO_2 + eTMP + fHUM + gSR + hB1 + iB2 + jB3 + kB4 + lB5 + mB6 + nB7 + oNDVI + pSAVI + qEVI + rLST + I \quad (5.8)$$

where a, b, c … , r are the coefficients of the regression model, and I is the intercept in the equation. The subset analysis reduces the number of input variables with the considered criteria (RSS, Adj. R^2, CP, BIC).

Once the input variables are selected, a PLS regression is applied to avoid the multicollinearity between the variable subsets. PLS is a technique applied in cases where traditional regression models fail, and the predictors have a high correlation, as shown in Equations 5.9 – 5.10.

$$X = TP^T + E \quad (5.9)$$

$$Y = UQ^T + \mathrm{F} \quad (5.10)$$

Where X is a *n x m* matrix of predictors, Y is a *n x p* matrix of responses; T and U are *n x l* matrices that are, respectively, projections of X and projections of Y; P and Q are, respectively *m x l* and *p x l* orthogonal loading matrices; and matrices E and F are the error terms. The decompositions of X and Y are made in order to maximise the covariance between T and U. Additionally, PLS generate an orthogonal transformation to obtain components by finding the most appropriate model to explain the variance starting from the maximise covariance matrixes [76]. In the case of remote sensing data, some studies consider multicollinearity when the same sensor is used to obtain different variables [147,193]. Finally, the validation is performed by cross-validation (Figure 5.3) and the criterion to accept or reject models where R^2, RMSE, predicated vs measured graphic and residuals analysis. The R-packages used were *pls* and *plsdepot*.

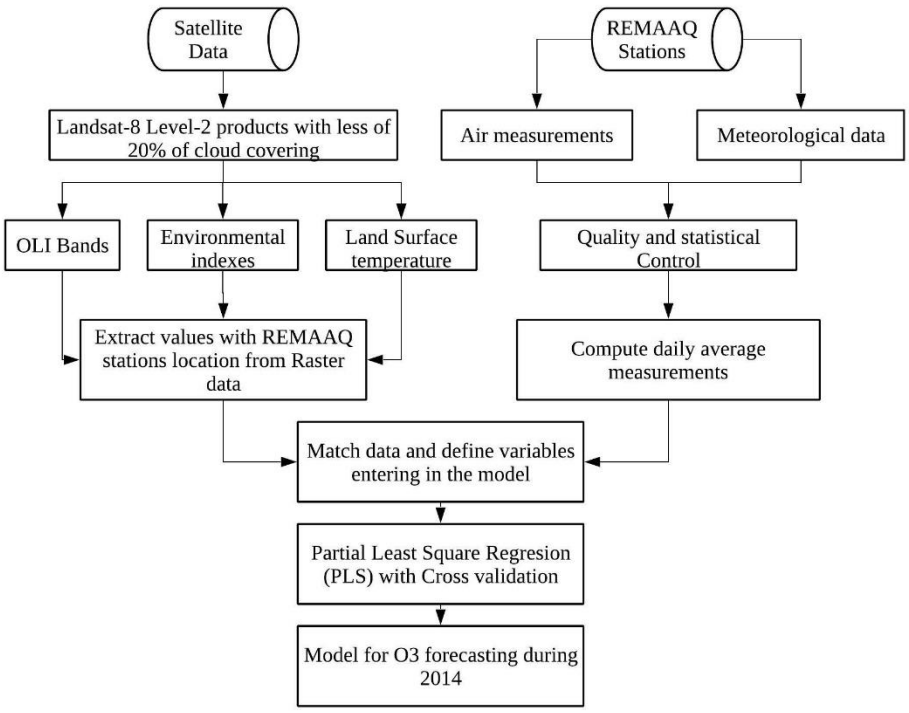


Figure 5.3. Methodology workflow

5.4 Results

5.4.1. Building the ozone LUR model

The LUR model tested 19 variables (18 independent variables or predictors and O_3 as the dependent variable), matching all variables (air measurement data, meteorological data and remote sensing data). The result is a database with 36 observations, where most of the remote sensing data variables show a high correlation (Figure 5.4). The high correlation or multicollinearity (in some cases near 1) indicates that some variables are highly related, such as NDVI, SAVI and EVI, or the visible bands (B1, B2, B3, B4). On the other hand, the highest correlation between all predictors with O_3 is PM2.5, showing a value of -0.44. The highest correlation considering only the remote sensing data variables is B6 with 0.22.

To find the model with the best fit, a stepwise regression subset is used. In the first instance (Figure 5.5), the coefficient of determination (R^2) is near 0.68, considering all 18 independent variables to build the model. The subset variables are analysed by the less Akaike information criterion (AIC) and the maximum Adj. R^2.

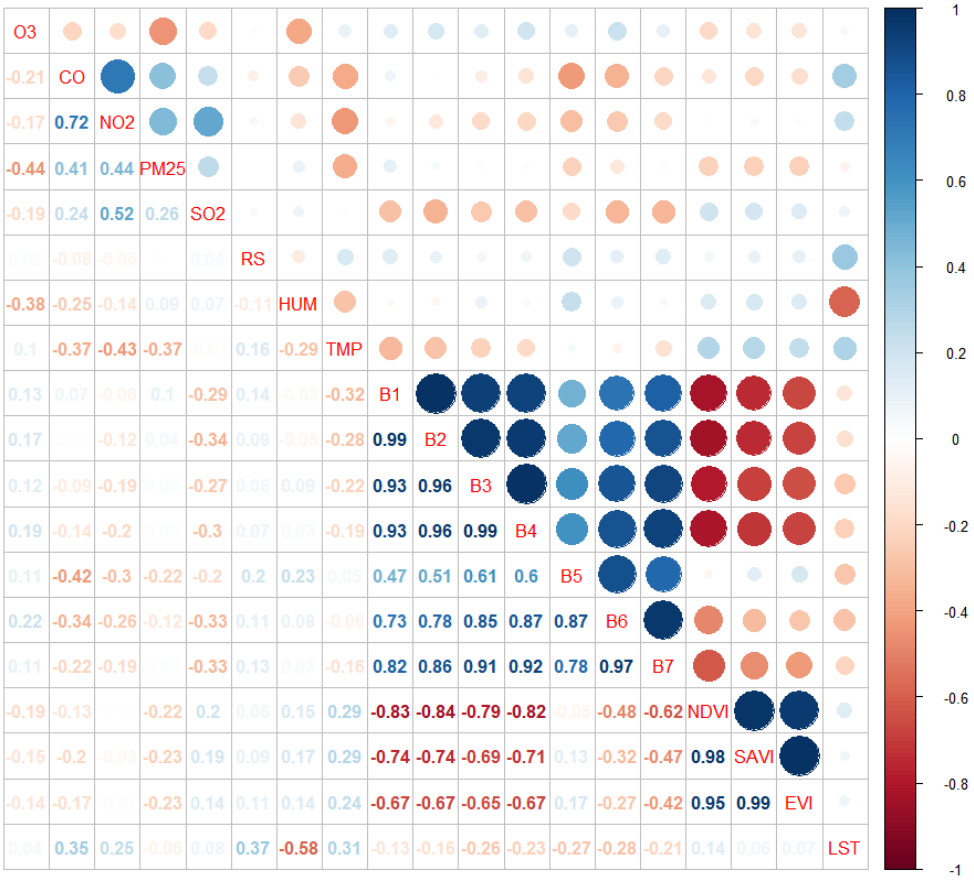


Figure 5.4. Correlation graph between input variables.

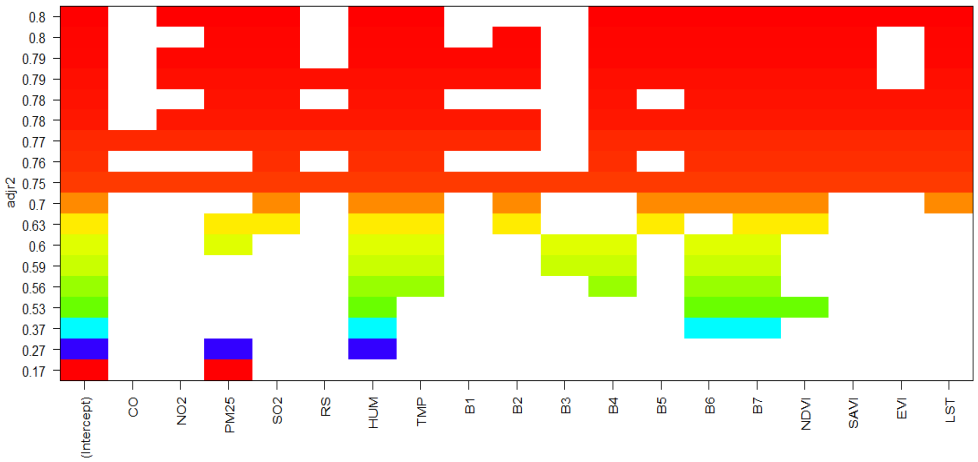


Figure 5.5. Variable combinations with their corresponding R^2 values as part of the subset task to select the model with the best fit.

The preliminary predictors are known (Figure 5.5), so to find a simple model with fewer input variables, a new subset of variables, applying RSS, Adj. R^2, CP and BIC criteria are analysed (Figure 5.6). Analysing the four criteria, eleven independent variables are used to build the simplest model (PM2.5, HUM, TMP, B2, B4, B5, B7, NDVI, SAVI, EVI).

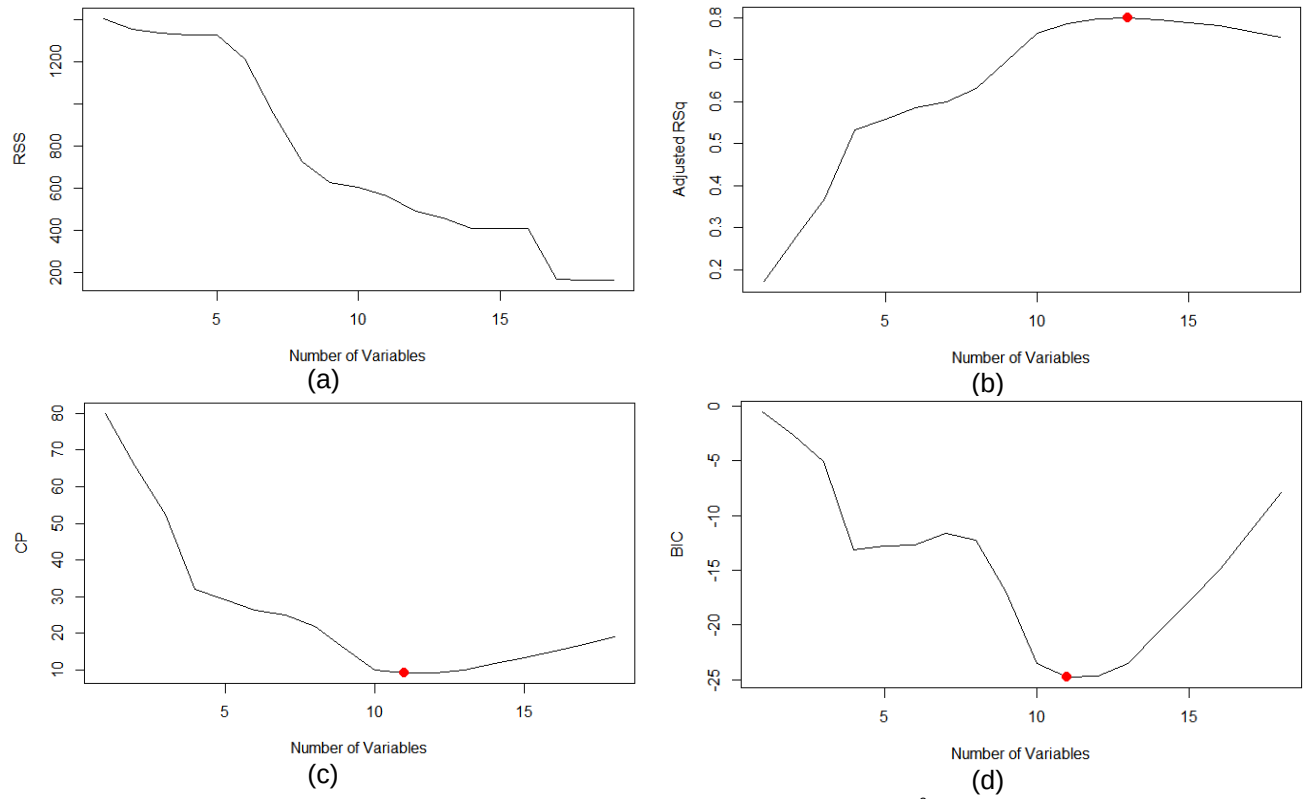
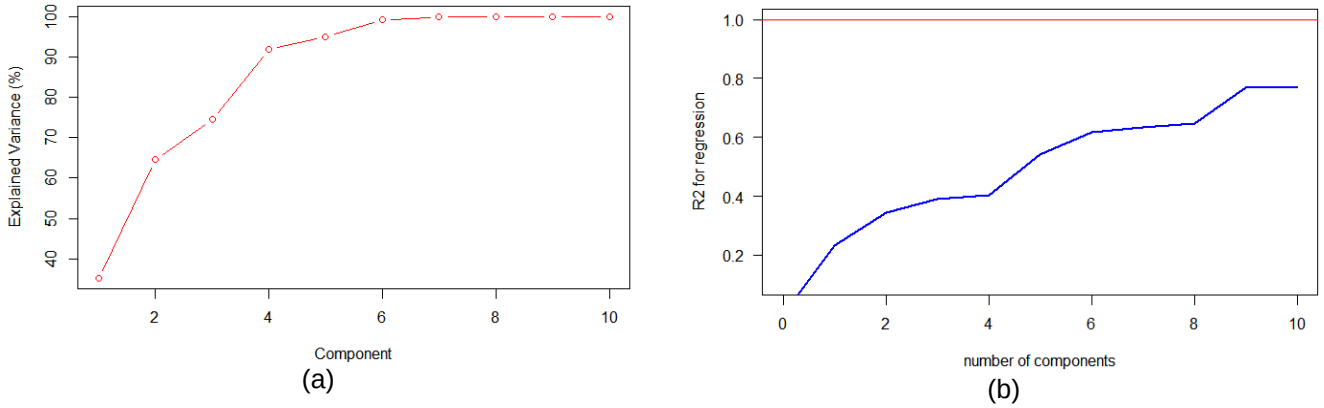


Figure 5.6. Subset analysis to select variables with different criteria: (a) RSS; (b) Adj. R^2; (c) CP; (d) BIC. The red point shows the optimal value of variables for each criterion.

The eleven variables chosen were then considered in the PLS analysis (Figure 5.7). The number of components in PLS regression was nine. These components explain most of the percentage of variance (Table 5.4), after cross validation (data not shown). The R^2 obtained was 0.77, and the RMSE was 3.03 through the PLS regression.



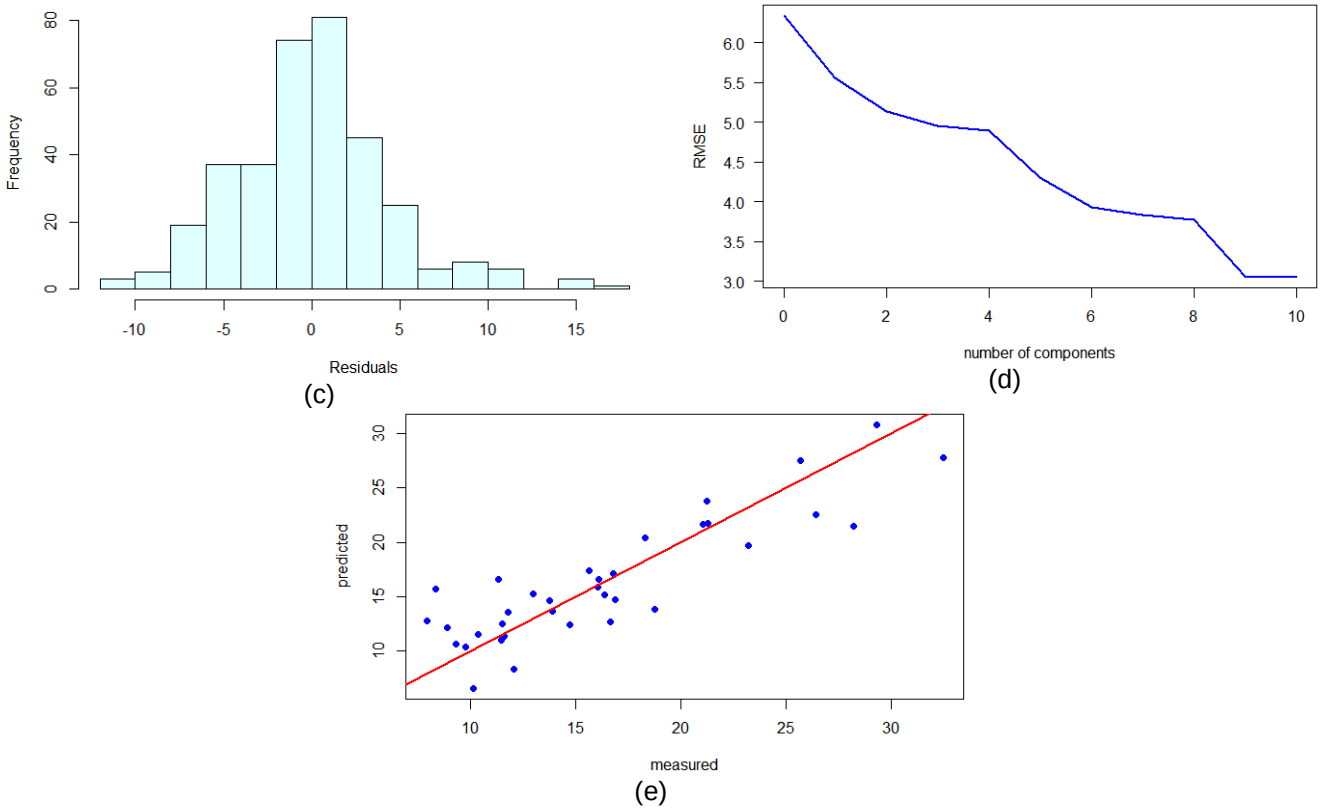


Figure 5.7. PLS analysis a) The number of components that explain the variance. b) The number of components to obtain the highest R^2. c) The histogram of the residuals. d) The number of components to obtain the lowest RMSE. e) Measured vs. predicted values with PLS regression.

Table 5.4. Variables explained variance by PLS components (t1, t2, …, t6). The red text shows the maximum variance explained with nine components, considering O_3 as the dependent variable.

Variable	t1	t2	t3	t4	t5	t6	t7	t8	t9
PM2.5	0.148	0.655	0.660	0.787	0.897	0.999	1.000	1.000	1.000
HUM	0.212	0.433	0.442	0.593	0.775	1.000	1.000	1.000	1.000
TMP	0.017	0.350	0.902	0.978	0.979	1.000	1.000	1.000	1.000
B2	0.611	0.918	0.918	0.947	0.955	0.966	0.998	1.000	1.000
B4	0.609	0.934	0.948	0.994	0.995	0.998	0.998	1.000	1.000
B5	0.123	0.158	0.362	0.994	0.997	0.999	1.000	1.000	1.000
B7	0.460	0.714	0.777	0.951	0.952	0.974	0.996	1.000	1.000
NDVI	0.515	0.873	0.904	0.987	0.987	0.993	0.994	1.000	1.000
SAVI	0.435	0.740	0.805	0.989	0.990	1.000	1.000	1.000	1.000
EVI	0.387	0.677	0.729	0.957	0.958	0.991	0.999	1.000	1.000
R^2	0.232	0.345	0.390	0.404	0.541	0.617	0.634	0.646	0.768

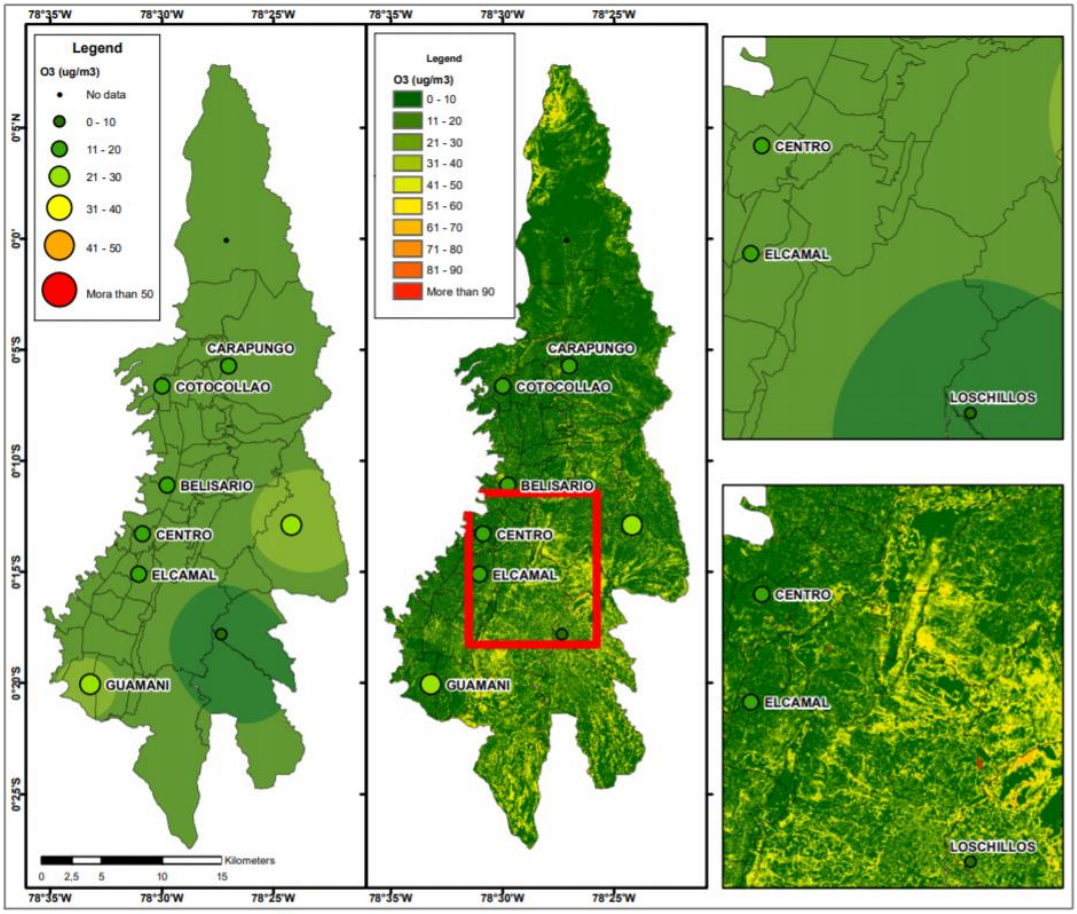
Avoiding the multicollinearity, the PLS regression is applied, presenting values different from 1 in the correlation matrix between the variables and the components (Table 5.5). Moreover, cross-validation is applied to the components. Equation 5.9 shows the resulting model to retrieve O_3 during 2014, considering the dataset.

$$O_3 = -0.47PM2.5 - 3.41TMP - 0.34HUM - 1371.47B2 + 9449.41B4 - 7852.43B5 - 436.68B7 - 1028.50NDVI + 4961.14SAVI + 1178.61EVI + 66.06 \quad (5.9)$$

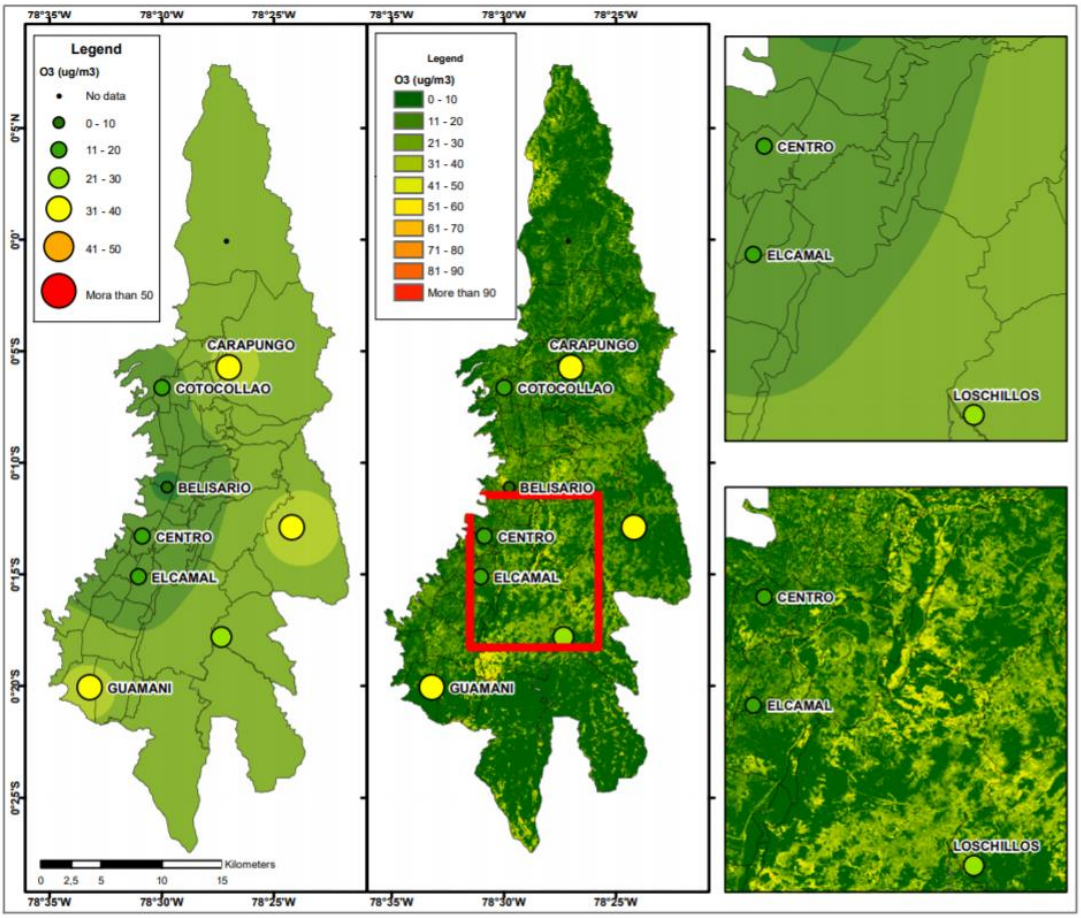
Table 5.5 Correlation matrix between the variables and the PLS components.

Variable	t1	t2	t3	t4	t5	t6	t7	t8	t9
PM2.5	-0.38514	-0.71215	-0.06697	-0.35607	0.33247	-0.31854	0.03311	-0.01169	0.00006
HUM	-0.46074	-0.46963	-0.09577	0.38916	-0.42616	0.47404	-0.01179	0.00889	-0.00005
TMP	0.13159	0.57670	-0.74288	-0.27615	-0.02596	0.14489	0.01716	0.00092	-0.00003
B2	0.78187	-0.55363	0.01333	0.17146	-0.08946	-0.10447	0.17864	-0.04175	-0.00072
B4	0.78063	-0.56987	-0.11695	0.21595	-0.01183	0.05406	0.00119	-0.04905	0.00434
B5	0.35068	-0.18755	-0.45108	0.79548	0.05186	-0.04768	0.01853	-0.01613	-0.00254
B7	0.67796	-0.50463	-0.25075	0.41702	0.02133	-0.14883	-0.14855	0.06528	-0.00017
NDVI	-0.71749	0.59850	-0.17601	0.28787	-0.02397	-0.07591	-0.02313	-0.07924	-0.00011
SAVI	-0.65920	0.55261	-0.25518	0.42854	0.03350	-0.09995	0.00418	0.00529	0.00143
EVI	-0.62209	0.53862	-0.22889	0.47707	0.02165	-0.18231	0.08867	0.03466	0.00153
O3	0.48204	0.33513	0.21357	0.11803	0.36972	0.27520	0.13251	0.10736	0.34908

Finally, Equation 5.9 allows mapping the O_3 concentration during 2014 (Figure 8).



(a)



(b)

Figure 5.8. Maps of O_3 during 2014: (a); January; (b) July maps obtained from Equation 8. The left map is with an inverse distance weighting (IDW) technique while the centre map is applying the O_3 model in all the study area. The right maps are a zoom in an assessment area (red square).

5.5 Discussion

The main goal of this study was to establish a model to retrieve O_3 from several input variables, implementing a variant of the classical LUR model. In most cases, LUR models are used to model air pollutants from road networks, land use, building density, MODIS AOD, population density and other geographic variables [132,143,162,194–196]. In this study, the variables selected are air pollution measurements, meteorological data (MD) and remote sensing data. The air pollution measurements and MD were obtained from REMMAQ stations. Moreover, considering the accuracy of LUR models in order to retrieve air pollutants (R^2 values between 0.45 and 0.80) [132,143,162,194–196], ten Landsat-8 images were selected to retrieve O_3 in Quito-Ecuador. Most LUR models use MODIS data. However, MODIS data probably do not have the accuracy and the quality to model pollutants or other environmental variables in middle cities [167].

To select the predictor variables, a subset was applied considering 19 variables (18 independent variables and O_3 as the dependent variable), obtained a preliminary best fit model with the 18 variables (R^2=0.68). However, to find the best fit and simplest (with

the lowest number of predictors) model, four criteria (RSS, Adj.R^2, CP, BIC) are analysed, resulting in a model with ten independent variables (PM2.5, HUM, TMP, B2, B4, B5, B7, NDVI, SAVI, EVI), showing an R^2 of 0.72 considering stepwise regression. In most of the subsets, the remote sensing data variables B1, B2, B6 and B7 appear, showing the relation between these bands with O_3. Thus, B1 and B2 reflect the blues and violets related to the aerosol presence [10]. Additionally, B6 and B7 reflect the short infrared related to greenhouse gas absorption [197]. Some studies that use LUR models employed stepwise regression to automatically find the predictors in a model [26,141]. However, the main problem with stepwise regression is not allowing a multicollinearity analysis [74]. PLS regression is used in some studies to compute the LUR model [163,195] to avoid multicollinearity. PLS builds a model with latent variables (components) as independent variables [76]. Moreover, PLS regression is used when we have a model with few observations [198]. If a high correlation is present between variables, a PLS regression is used to build the model, where nine components explain most of the variance and obtained an R^2 value of 0.768. This value is higher than R^2 in the stepwise regression (R^2 = 0.72) and avoids the multicollinearity of remote sensing variables.

The final model can be mapped, in comparison with other techniques, such as thematic point maps, interpolation or geostatistical analysis (Figure 8), showing a robust perception of spatial concentration of O_3 in the city, and these maps can be used as input to make a more accurate air pollution analysis.

The limitation is the few observations used to build the model because our model requires some data from the REMMAQ stations and sometimes these data are incomplete or unavailable. On the other hand, the remote sensing variables depend on the number of clouds. Quito is known as a city with a high cloud density during the year [152], and this factor limits the computation of LUR models. A possible alternative can be to combine different sensors with high spatial and temporal resolution and use similar techniques to PLS to compute the model.

Another limitation is the generation of a raster to each independent variable. In the case of remote sensing, data are not a problem considering all images over the study area, but the air pollutant measurements and MD raster can be limited. They were obtained with a geostatistical technique as inverse distance weighting (IDW) [199]. Nevertheless, this kind of technique works fine in a region with some stations, but in Quito, we only have nine stations (Figure 5.8). Therefore, in future work, we will propose the use of only remote sensing data to spatialize air pollutants in Quito.

5.6 Conclusion

A spatial estimation was performed in Quito to obtain the O_3 spatial concentration in 2014. The spatial estimation was computed by a variant of LUR models with PLS regression. LUR models can explain the spatial concentration of an air pollutant, helping in urban planning, environmental analysis and governmental decisions. Moreover, the idea of having a variant of LUR models with variables from remote sensing sensors different from MODIS will help to build more accurate models. The main limitation is related to the small quantities of field data available. In future work, we will try to find new alternatives only considering the use of remote sensing data as input without other field data variables.

6. Article 4: Spatial Modeling of Chronic Respiratory Diseases Based on Machine Learning Techniques—An Approach Using Remote Sensing Data and Environmental Variables.

Cesar I. Alvarez-Mendoza[1,2*], Ana Teodoro[1,3], Alberto Freitas[4] and Joao Fonseca[4]

[1] University of Porto, Department of Geosciences, Environment and Land Planning, Faculty of Sciences, Rua Campo Alegre 687, Porto 4169-007, Portugal; up201510599@fc.up.pt

[2] Universidad Politécnica Salesiana, Grupo de Investigación Ambiental en el Desarrollo Sustentable GIADES, Carrera de Ingeniería Ambiental, Quito 170702, Ecuador; calvarezm@ups.edu.ec

[3] Earth Sciences Institute (ICT), Pole of the FCUP, University of Porto, Porto 4169-007, Portugal; amteodor@fc.up.pt

[4] Department of Community Medicine, Information and Health Decision Sciences (MEDCIDS), Faculty of Medicine, University of Porto, Rua Dr. Plácido da Costa, 4200-450 Porto, Portugal; alberto@med.up.pt, jfonseca@med.up.pt

[5] CINTESIS - Center for Health Technology and Services Research, Faculty of Medicine, University of Porto, Rua Dr. Plácido da Costa, 4200-450 Porto, Portugal

Received: 25 July 2019 (under review)

Journal: **International Journal of Environmental Research and Public Health. MDPI. Special Issue Innovations in Remote Sensing and GIS for Environmental, Urban and Public Health Sciences**

6.1 Abstract

Over the last few years, the use of remote sensing data to retrieve air pollution variables through land use regression (LUR) models has become very popular. LUR models are an effective alternative to predict air quality, and some studies have established a possible relationship between environmental variables and respiratory health parameters. This study proposes that there is a relationship between remote sensing data (Landsat 8) and environmental variables (air pollution and meteorological data) that can be used to determine the number of hospital discharges of patients with chronic respiratory diseases in Quito, Ecuador, between 2013 and 2017. The main objective of this study is to establish and evaluate an alternative LUR model that is capable of calculating the prevalence of chronic respiratory diseases, in contrast with traditional LUR models, which typically assess air pollutants. Moreover, this study also evaluates different analytic techniques (multiple linear regression, multilayer perceptron, support vector regression, and random forest regression) that often form the basis of spatial models. The results show that machine learning techniques, such as support vector machine, are the most effective in computing such models, presenting the lowest root-mean-square error (RMSE). Additionally, in this study, we show that the most significant remote sensing predictors are the blue and infrared bands. Our proposed model is a spatial modeling approach that is capable of determining the prevalence of chronic respiratory diseases in the city of Quito, which can serve as a useful tool for health authorities in policy- and decision-making.

Keywords: remote sensing; machine learning; respiratory disease; spatial models; Quito

6.2 Introduction

During the last few years, remote sensing data have increasing been used in monitoring, spatial predictive modeling, surveillance, and risk assessment with respect to human health [23]. These human health studies have also been associated with air pollution spatial modeling, which is connected to some vector-borne [61] and respiratory diseases [26]. In this context, spatial models relying on remote sensing data have been developed to identify different air pollutants, the most common of which are particulate matter (PM) [6,129], nitrogen dioxide (NO_2) [200], tropospheric ozone (O_3) [34], sulfur dioxide (SO_2) [201], and carbon dioxide (CO_2) [13]. The aforementioned air pollutants are greenhouse gases and precursors of global warming [202]. Moreover, evidence of the adverse effects of exposure to air pollutants (PM2.5, PM10, O_3) on health has been collected in several

countries around the world [203]. Specifically, air pollution is a threat to respiratory health, and several chronic respiratory diseases (CRDs), such as asthma, chronic obstructive pulmonary disease (COPD), and others, represent nearly 6% of global annual deaths [1,2]. According to the World Health Organization (WHO), 92% of people around the world live in places with poor outdoor air quality, where the main risk factors of developing a CRD are related to the climate and the environment [1,4].

One of the most famous missions in satellite remote sensing is the Landsat program, launched in 1972. The most recent program satellite is the Landsat 8, which has the Operational Land Imager (OLI) and Thermal Infrared Sensor (TIRS) [204] on board. This satellite provides a wide spatial–temporal perspective of the Earth, enabling a variety of applications and retrieving several variables, such as vegetation, land use, aerosol particles, and environmental and meteorological information, which can be retrieved and analyzed [10]. Due to the potential of the variables collected by remote sensing of the Earth's environment [40], it is possible to develop models to analyze air pollutants. Such models typically use air quality monitoring network (AQMN) data and remote sensing variables to conduct spatial modeling of air pollutants, using remote sensing-derived parameters in the form of environmental indexes, such as the normalized difference vegetation index (NDVI) [205], and measures, such as aerosol optical thickness (AOT) [19,190]. It is important to note that sensors, such as the MODIS instruments on Terra and Aqua and Landsat 8's OLI, allow us to obtain (directly or indirectly) this information. The Terra/Aqua MODIS instruments have an AOT product with a low spatial resolution (3 x 3 km), which is ideal for regional studies [166]. Landsat 8's OLI is capable of retrieving AOT in fine spatial resolution (30 meters); however, AOT information from ground stations is also needed [142]. AOT measurements are retrieved by the blue and red bands of Landsat 8's OLI [50]; its infrared bands are also used to retrieve O_3 measurements [181,206].

With respect to health studies based on remote sensing data, predictive models have been used to analyze air pollutants by combining geographic variables (traffic, land use, population, etc.) with remote sensing data, in the form of land use regression (LUR) models [159,168]. Thus, a LUR model could potentially be used to investigate the possible relationship between hospital discharge rates and certain environmental variables [26,207]. A hospital discharge is defined as the release of a patient who has stayed at least one night in the hospital, including people who die in hospital care [208]. However, most LUR models do not consider the dynamics of geographical variables, because such variables are sometimes out of date or obsolete [209]. Some health studies have related hospital discharge with exposure to different traffic-related pollution,

in which the NDVI and MODIS AOT are the most commonly used predictors. These studies aim to find a possible relationship between air pollution exposure and hospital discharge [210,211].

LUR models use analytic techniques, such as multiple linear regression (MLR), stepwise regression (STW), and multiple logistic regression [212,213]. However, these techniques do not analyze the correlation between predictors, and it is well known that remote sensing variables have a high correlation or multi-collinearity [214]. An alternative to MLR is the use of more complex models, such as machine learning techniques (MLTs) in order to avoid multi-collinearity. Examples of non-linear MLTs are multilayer perceptron (MLP), support vector regression (SVR), and random forest regression (RFR), among others.

In this context, the aim of this study is to establish and compare spatial empirical models, based on LUR models, that are capable of determining the number of hospital discharges of patients with CRDs (HCRD) in Quito, Ecuador, between 2013 and 2017, using remote sensing data, air pollution field measurements, and meteorological data as predictors and considering three different complex machine learning techniques: MLP, SVR, and RFR. The spatial model selected will allow us to map the prevalence of HCRD. This approach will provide insight into and an understanding of the most significant spatial predictors and the spatial distribution of HCRD in the city of Quito. Furthermore, the present study is an innovative approach to the use of remote sensing data in human health studies.

6.3 Materials and Methods

6.3.1. Study Area

The study area is the most populated zone of Quito, Ecuador. The area is divided into 45 administrative urban districts. Its latitude is 0º30’S to 0º10’N, its longitude is 78º10’W to 78º40’W, and the equatorial line crosses through it (Figure 6.1). Quito’s annual median temperature is 17 °C, and its elevation is about 2800 meters above sea level. This specific study area was chosen for the following reasons: (i) it is covered by nine AQMN stations; (ii) its road traffic is relatively high; and (iii) the urban downtown is located in this area. The influence zones of each AQMN station were established through Thyssen polygons and their respective intersection with the urban districts.

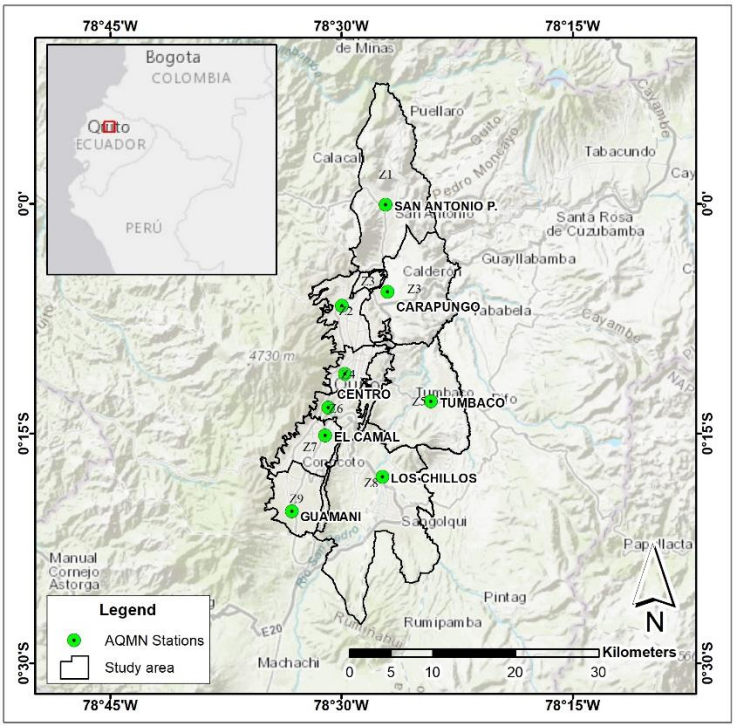


Figure 6.1. The study area (Quito, Ecuador). The green dots represent the air quality monitoring network (AQMN) stations and their influence areas.

6.3.2. Data Collection

6.3.2.1 Remote Sensing Data

Between 2013 and 2017, 46 Landsat 8 level 2 images were acquired over the study area. The on-demand images were obtained from the Land Satellite Data Systems (LSDS) Science Research and Development (LSRD) website (https://espa.cr.usgs.gov/). The main advantage of level 2 images is that they use the Landsat 8 Surface Reflectance Code (LaSRC) to generate products with geometrical, radiometric, and atmospheric corrections [51]. These products have a spatial resolution of 30 meters. The products used in this study as predictors are the surface reflectance (SR) OLI bands, the top of the atmosphere (TOA), brightness temperature (BT), and some pre-processed indexes, such as the NDVI [54], the soil-adjusted vegetation index (SAVI) [55], and the enhanced vegetation index (EVI) [56]. Moreover, considering the high cloud density in the Andean Region [32], the images were filtered, and only images with a maximum of 10% cloud density over the study area were considered.

BT was converted to land surface temperature (LST) using the emissivity equation according to [157,215] and the inversion of Planck's function, as shown in Equation (6.1):

$$LST = \frac{BT}{\left(1+\left(\frac{\lambda * BT}{\rho}\right) ln\varepsilon\right)} - 273.15 \quad (6.1)$$

where BT is obtained from Landsat 8 level 2 images in kelvin degree (K), λ is the center wavelength of the Landsat 8 TIR 1 band (10.8 μm) [156], ρ is expressed in Equation (6.2), and ε is the emissivity derived from Equation (6.3), which has to be selected according to the NDVI evaluation in the study area. The result is the LST in degrees Celsius (°C).

$$\rho = \frac{h*c}{s} \tag{6.2}$$

where h represents Planck's constant (6.63e-34 Js), c is the speed of light (2.99e-8 ms^{-1}), and s is the Boltzmann constant (1.38e-23 $m^2kgs^{-2}K^{-1}$).

$$\varepsilon = \begin{cases} \varepsilon_s, NDVI < NDVI_s \\ \varepsilon_s + (\varepsilon_v - \varepsilon_s)P_V, NDVI_s \leq NDVI \leq NDVI_v \\ \varepsilon_v, NDVI > NDVI_v \end{cases} \tag{6.3}$$

where ε_s is the emissivity for the soil (0.973) and ε_v is the emissivity for the vegetation (0.985) [157]. $NDVI_v$ is the NDVI for the vegetation (0.2), and $NDVI_s$ is the NDVI for the soil (0.5) [59]. P_v represents the proportion of vegetation in the study area according to Equation (6.4).

$$P_V = \left(\frac{NDVI - NDVI_s}{NDVI_v - NDVI_s}\right)^2 \tag{6.4}$$

Moreover, the cloud pixels are removed from each satellite image considering the information available in the level 2 pixel quality band (Band QA). All the processes were computed on RStudio with the raster v2.9-5 package.

6.3.2.2 Field Measurement Data

Most of the models that calculate air pollutants require airfield measurements. In this work, field data were obtained from the Quito AQMN, known as "Red Metropolitana de Monitoreo Atmosférico de Quito" (REMMAQ) [38]. This AQMN has been in operation since 2002, providing hourly field measurements of air pollutants and meteorological variables. REMMAQ has nine georeferenced stations (Figure 1), which collect the following air pollution variables of interest to this study: carbon oxide (CO), PM less than 2.5 and 10 microns (PM2.5 and PM10, respectively), SO_2, O_3, and NO_2. The following meteorological variables were considered in this study: pressure, wind direction, relative humidity, precipitation, wind speed, air temperature, and solar irradiance. The Environmental Secretary of Quito manages the REMMAQ, and the data are available to download for free on her website (http://www.quitoambiente.gob.ec/ambiente/index.php/datos-horarios-historicos).

Spatial air pollutant rasters for each trimester of every year were computed using the

inverse distance weight (IDW) algorithm [5]. All the information was processed with the R packages *rgal v1.4-4* and *gstat v2.0-2*.

6.3.2.3. Hospital Discharges of Patients with Chronic Respiratory Diseases

The National Institute of Statistics and Census (INEC) is the official government institution in Ecuador in charge of collecting and disseminating information about population and other socioeconomic statistics and variables. This information is public and available on a district scale (http://www.ecuadorencifras.gob.ec/camas-y-egresos-hospitalarios/). One of the variables included in this information is the number of hospital discharges (the number of released patients who stayed at least one night in the hospital, including people who died in hospital care) organized by their home district. This variable is classified according to the International Classification of Diseases 10th version (ICD-10) from the WHO [68]. Considering the aim of this study, only hospital discharges of patients with CRDs were considered—those with ICD-10 classification codes of J40–J47. This filter includes the most significant CRDs, such as asthma and bronchitis. A summary of hospital discharges in each AQMN area of influence was computed for each trimester of each year. The main reason to group the dataset by trimester was the availability of matched data. Furthermore, population data are necessary to compute HCRD (the number of patients per 10,000 people who are admitted to the hospital with a CRD) to compare the different urban districts. This variable is a continuous dependent variable.

6.3.3. Input Dataset

In order to compile a unique dataset encompassing the remote sensing data, environmental variables (air pollution and meteorological field data), and HCRD, all the variables were correlated by trimester, year, and AQMN area of influence. Clipping the Shapefile of the AQMN area of influence and the variables allowed us to obtain the median trimestral variable for each AQMN area of influence. Table 6.1 shows the variables used in this study and their respective statistics.

Table 6.1. Descriptive statistics of the input variables

No.	Variable	Min.	Max.	Mean	Median	Units/scale
0	HCRD	0.334	23.433	4.463	3.689	Hospital discharges per 10,000 people with chronic respiratory disease
1	Coastal aerosol band (B1)	0.029	0.077	0.056	0.060	reflectance (0–1)

2	Blue band (B2)	0.034	0.095	0.068	0.072	reflectance (0–1)
3	Green band (B3)	0.062	0.136	0.098	0.101	reflectance (0–1)
4	Red band (B4)	0.050	0.149	0.105	0.111	reflectance (0–1)
5	Near–infrared (NIR) (B5)	0.182	0.291	0.231	0.228	reflectance (0–1)
6	Short-wave infrared 1 (SWIR 1) (B6)	0.170	0.268	0.208	0.206	reflectance (0–1)
7	Short-wave infrared 2 (SWIR 2) (B7)	0.092	0.218	0.159	0.163	reflectance (0–1)
8	Normalized Difference Vegetation Index (NDVI)	0.171	0.721	0.359	0.312	0-1
9	Soil-Adjusted Vegetation Index (SAVI)	0.101	0.408	0.209	0.184	0-1
10	Enhanced vegetation index (EVI)	0.106	0.428	0.217	0.190	0-1
11	Land Surface temperature (LST)	15.031	39.758	26.232	26.299	degrees Celsius
12	Pressure (P)	712.945	761.178	740.476	741.018	mb
13	Wind direction (WD)	58.155	273.426	142.357	146.345	degrees
14	Relative humidity (RH)	49.140	84.582	69.190	72.632	percentage (%)
15	Precipitation (PR)	0.000	4.443	0.406	0.000	mm
16	Wind speed (WS)	0.879	2.482	1.686	1.743	m/s
17	Air temperature (AT)	11.749	17.421	14.957	15.041	degrees Celsius
18	Solar irradiance (SR)	0.092	278.691	166.728	215.724	W/m^2
19	CO	0.435	0.852	0.622	0.598	$\mu g/m^3$
20	NO_2	11.458	35.256	23.055	22.169	$\mu g/m^3$
21	O_3	7.518	44.055	22.786	22.130	$\mu g/m^3$
22	PM2.5	10.441	23.504	16.490	16.316	$\mu g/m^3$
23	PM10	0.030	87.590	35.770	38.364	$\mu g/m^3$
24	SO_2	0.839	7.829	3.459	3.273	$\mu g/m^3$

6.3.3. Model Establishment

An LUR model is an empirical model that considers some geographical predictors as independent variables and a dependent variable. The first step in establishing such a model is the selection of the input predictors. The simplest model with the least number of independent variables should be found in order to avoid overfitting. Here, the Bayesian

information criterion (BIC) was considered to conduct backward elimination, by which the lowest BIC values were used to choose the predictors [216,217]. Then, the models were computed, considering different MLTs in order to compare linear (MLR) and non-linear regression models (MLP, SVR, and RFR). In each model, 80% of the dataset was used as training data, and 20% of the dataset was used as test data.

MLR is probably the simplest and most common analytic technique used in building a predictive model. It computes a linear relationship between the independent (predictors) and the dependent variables [218]. However, MLR does not analyze the correlation between predictors—a major limiting factor when considering remote sensing variables [147], which are highly correlated. In contrast, MLP with a back-propagation learning process is classified as an artificial neural network (ANN) model, and it can be used in the classification of remote sensing data. MLP uses a series of neuronal activities where the ideal is to have interconnection weights in a multilayer perceptron [77]. In this study, a non-linear MLP with an architecture defined by a hidden layer and six hidden nodes was computed according to [161] and evaluated. The R package *neuralnet v1.44.2* was used to compute the MLR. SVR is a non-linear transformation of an MLT, which works as a support vector machine (SVM) classifier. SVM and SVR work in a higher dimensional space. The main difference is that SRV uses a continuous number as a dependent variable [82]. The R package used to compute SVR was *e1071 v1.7-2*. Finally, the last MLT employed was RFR. It is based on ensemble learning, which uses the training dataset to generate multiple decision trees, making it less sensitive to the overfitting problem. The decision trees are simply combined according to their weights. Moreover, RFS is considered to be one of the most effective non-parametric ensemble learning methods in image analysis [85]. The R package *randomForest v4.6-14* was used to implement RFS in this study.

In the model evaluation, the coefficient of determination (R^2) between the observed values and the predicted values and the root-mean-square error (RMSE) were compared. Models (considering the test dataset) with a higher R^2 and lower RMSE were selected to develop a spatial map of HCRD for each trimester for each year. The final model developed a raster file with 30 meters of spatial resolution. Figure 6.2 shows the workflow of the methodology used in this study.

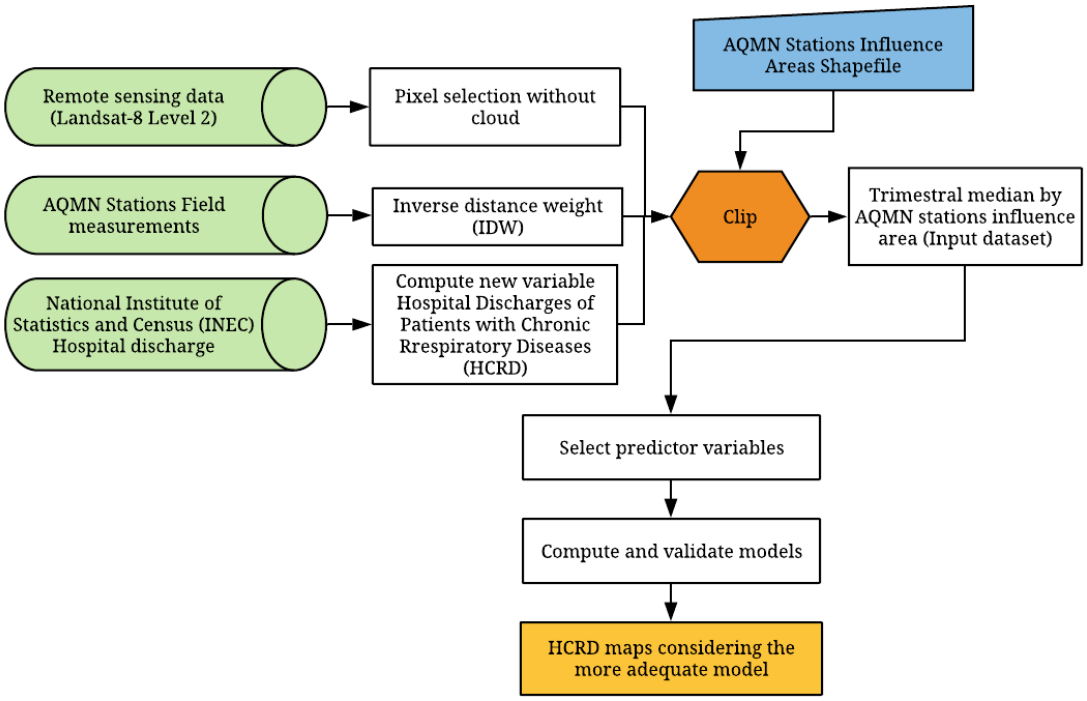


Figure 6.2. Workflow of the methodology applied in this work

6.4 Results

6.4.1. Selected Predictor Variables

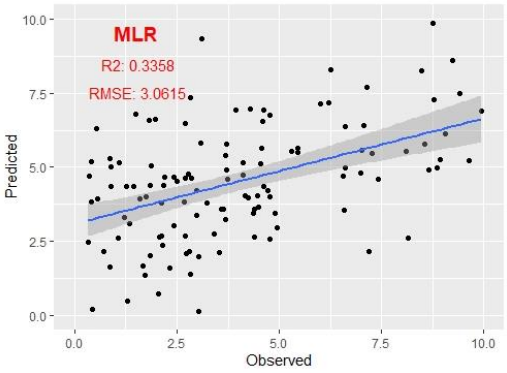
The final dataset considered 162 observations, which included all the variables (the remote sensing, environmental, and HCRD variables). The dataset consisted of 25 variables (one dependent variable and 24 predictors), including registers by trimester, year, and the AQMN area of influence. The lowest BIC values were chosen in order to consider only the most significant variables, avoiding multi-collinearity. A total of 10 predictors (B1, B2, B7, EVI, LST, RH, SR, AT, CO, and SO_2) were considered as inputs in all the MLTs (p-value < 0.050). Equation (6.5) shows the MLR established with the 10 predictors considered:

$$HCRD = I + aB1 + bB2 + cB7 + dEVI + eLST + fRH + gSR + hAT + iCO + jSO2 \quad (6.5)$$

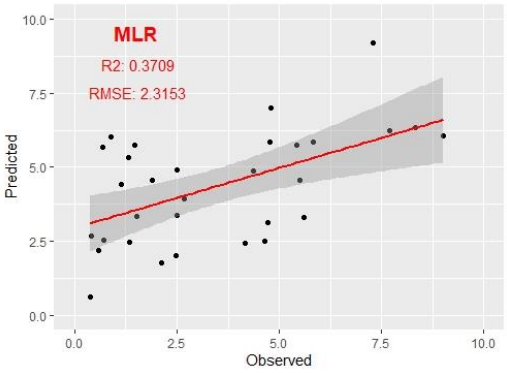
where HCRD is the hospital discharges per 10,000 people with chronic respiratory disease; I is the intercept; a, b, c, d, and e are the coefficients in each predictor; and the other variables are described in Table 6.1.

6.4.2. Comparison and Evaluation of the Models

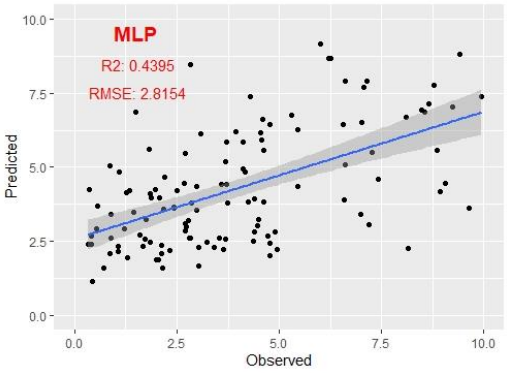
The results presented in Figure 6.3 allow us to analyze the relationship between the observed data and the predicted data considering the value of R^2.



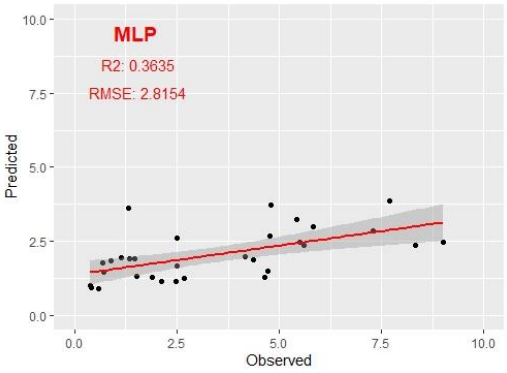
(a)



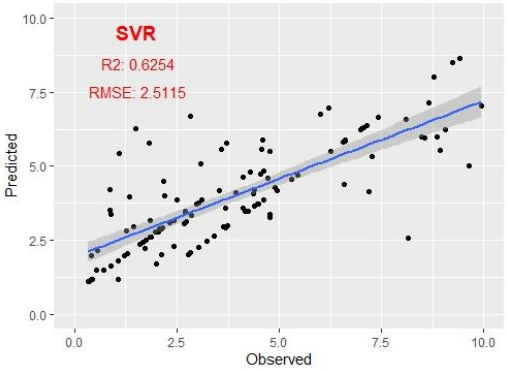
(b)



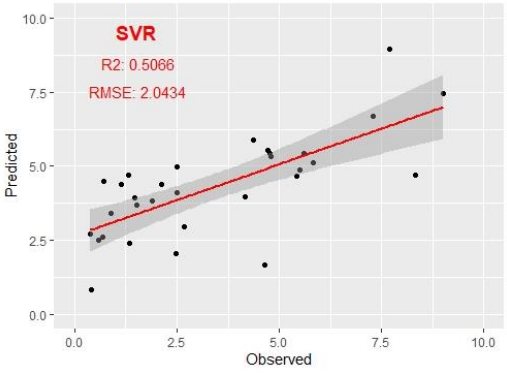
(c)



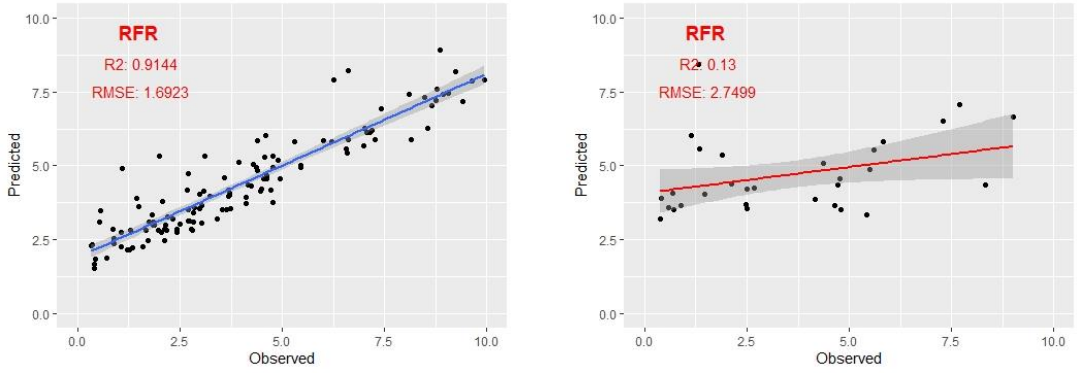
(d)



(e)



(f)

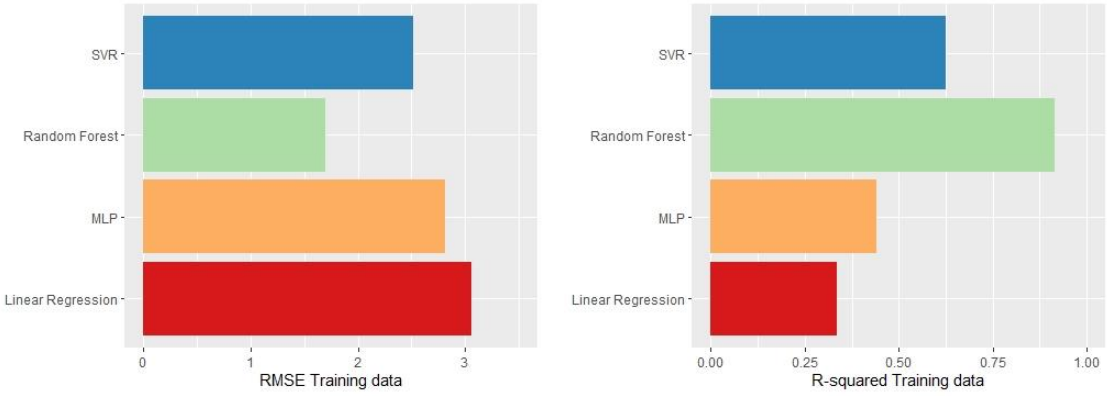


(g) (h)

Figure 6.3. Scatter plots of the different methods employed the model. The blue line represents the training data (a),(c),(e),(g), and the red line represents the test data (b),(d),(f),(h).

Figure 6.4 shows the comparison between R^2 and RMSE for all the models established. The non-linear models RFR and SVR showed the best adjustment both in the training data and in the test data.

According to the results presented in Table 6.2, the model with the lowest RMSE (1.6923) and the highest R^2 (0.9144), considering the training data, was the RFR. On the other hand, the model with the lowest RMSE (2.0439) and the highest R^2 (0.5066), considering the test data, was the SVR. The SVR model, considering the test data, was used to map HCRD. Figure 6.5 presents the SVR model by trimester (and year).



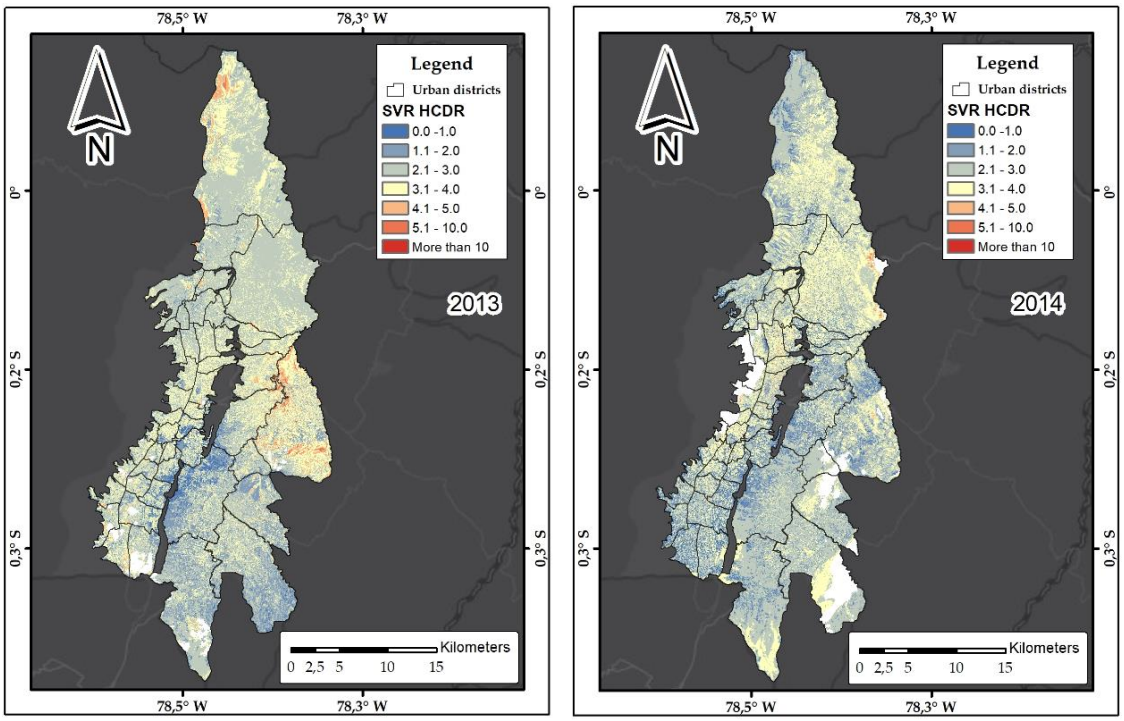
(a) (b)

(c) (d)

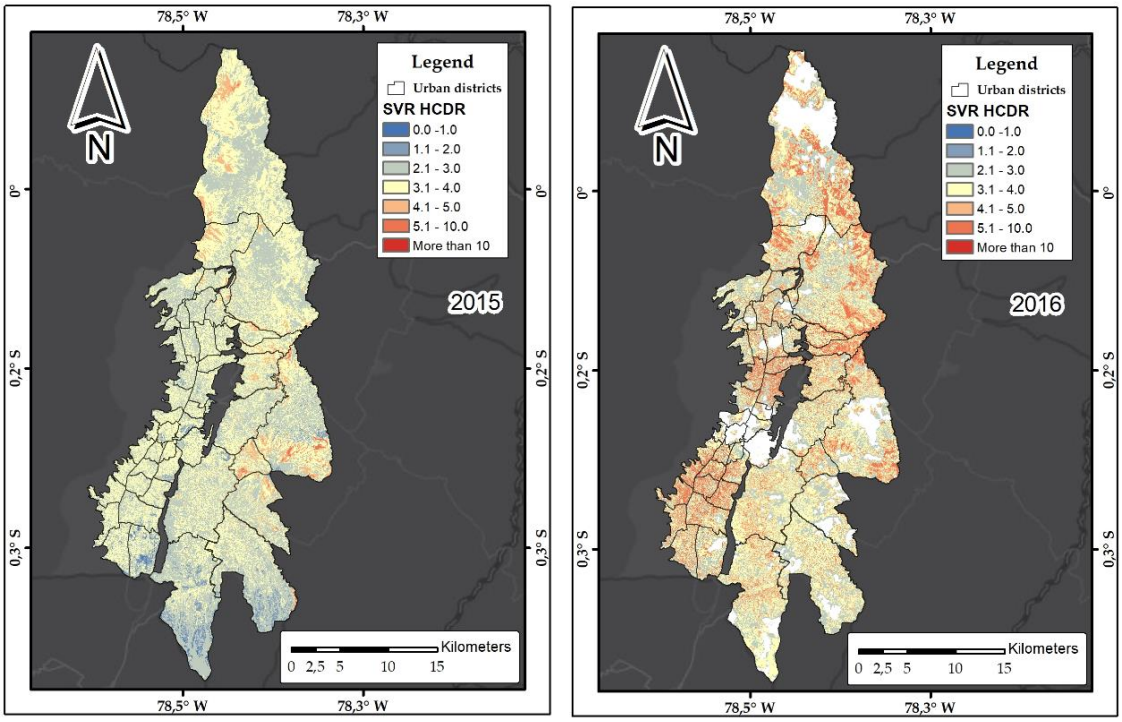
Figure 6.4. Comparison between models considering the RMSE and R^2: (a) RMSE training data; (b) R^2 training data; (c) RMSE test data; (d) R^2 test data.

Table 6.2. RMSE and R^2 for all the models tested.

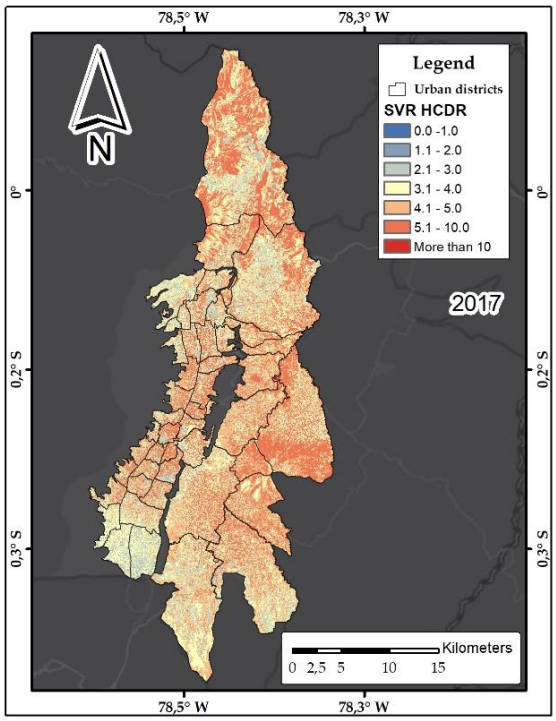
Model	RMSE Training Data	R^2 Training Data	RMSE Test Data	R^2 Test Data
Multiple Linear Regression (MLR)	3.0615	0.3358	2.3153	0.3709
Multilayer Perceptron (MLP)	2.8154	0.4395	2.7904	0.3635
Support Vector Regression (SVR)	2.5115	0.6254	2.0434	0.5066
Random Forest Regression (RFR)	1.6923	0.9144	2.7499	0.1300



(a) (b)



(c) (d)



(e)

Figure 6.5. HCRD maps considering the third trimester of the year (July–September) using the SVR model in (a) 2013, (b) 2014, (c) 2015, (d) 2016, and (e) 2017. The white areas show cloud presence.

6.5 Discussion

In most cases, LUR models are used to estimate air pollutants [34,36,136,219], and geographical variables, such as roads, traffic, land use, etc., are used to establish MLR models. However, these models include geographic variables that are not always available or updated in a timely manner. Some studies have also compared the air pollution calculated by LUR models with health data [16,26]. However, this comparison is only performed considering categorical variables [220] and not with numerical variables in order to quantify the value.

In this study, spatial models were developed to compute HCRD, considering remote sensing and environmental variables (air pollution and meteorological ground data) as predictors. The predictors were chosen considering their relationships with variables that may potentially affect respiratory health, such as vegetation, land use, climate, and air pollution. Air pollution is defined as the presence of one or more harmful substances in the air [221]. Some studies have shown that air pollution is a serious issue, posing a grave threat to respiratory health [14]. On the other hand, climate variables, specifically meteorological variables, such as temperature or humidity, have a direct influence on the potential to acquire CRDs [222,223]. Moreover, using RS data, it is possible to derive some of these environmental variables. One such example is the high correlation between the Landsat 8 blue and red bands with AOT [50]. As noted above, AOT influences the retrieval of air pollutants [18,20,224]. In this context, in this work, 24 predictors were considered as inputs in the original model with a matched dataset of 162 observations. Considering BIC, 10 significant predictor variables were selected for use in the final spatial HCRD models. The RS variable predictors included the coastal aerosol, blue, and SWIR-2 bands (bands 1, 2, and 7, respectively). The blue band is more correlated with PM [50] and the SWIR-2 band with O_3 concentrations [181,206]. Additionally, EVI and LST were also selected. The environmental variable predictors were CO and SO_2, and the meteorological variables were HR, SR, and AT. Several studies have already reported a correlation between these variables and the presence of CRDs [124].

Four MLTs were selected to compute the model: (i) linear MLR; (ii) non-linear MLP; (iii) SVR and; (iv) RFR. During the computation, 80% of the dataset was used as training data, and 20% of the dataset was used as test data. The main advantage of non-linear MLTs is the avoidance of concerns regarding multicollinearity. Several studies have found that the use of MLP and SVR with RS data improves the performance of regression models using ground true data [225]. RFRs are often implemented in prediction analyses, because they provide better accuracy [226]. The results show that the use of

RFR and SVR created the most successful models. SVR had the highest R^2 (0.5066) and the lowest RMSE (2.0434) with the test data. Thus, the SVR model helped to develop a spatial map of HCRD in different trimesters between 2013 and 2017 (Figure 6.5). The third trimester was selected because there were more available images during that trimester during the five years of the study period. Additionally, there was more variation in the rates of hospital discharges in September, according to the input data. It is also worth noting that there was a significant increase in reported air pollutant concentrations in some areas between 2013 and 2017 [170]. Indeed, in the north and east regions of the city, there were higher values of HCRD, which was likely due to the fact that these areas had higher rates of air pollutants, traffic, and population. Thus, these results allow us to identify possible trends in the growth patterns of CRDs in the next few years.

The main limitations of this study were as follows: (i) There were a limited number of satellite images available without high cloud density [32]. In future research, a possible improvement could be the use of more sensors to combine data or to develop and apply new techniques to remove cloud interference [152]; (ii) The REEMAQ and INEC data were incomplete for some months during all study years. In some cases, the stations were unavailable or did not have complete quality data. On the other hand, some hospital discharge data were lacking information regarding location, or such information suffered from poor-quality codification or registration. We discarded these data in order to obtain a more accurate dataset; however, in future work, we will extend this study for a longer period of time in order to improve our models; (iii) The percentage of training and test data may not have been ideal. Our future research could consider different cutoff values in the dataset; (iv) Despite the fact that the spatial HCRD maps give us a general idea of the presence of CRDs and possible future trends, these maps must be improved with more input data.

In this context, the models presented in this work, despite having some limitations, were shown to be valid tools in the prediction of HCRD, which will provide local health authorities with valuable information to improve policy- and decision-making.

6.6 Conclusions

This study proposed an innovative, alternative use of LUR models to establish a spatial modeling approach to calculating the number of hospital discharges of patients with CRDs in Quito, Ecuador. The proposed model considered geographical predictors, specifically RS data (Landsat 8) and environmental variables (air pollution and meteorological information) from 2013 to 2017. The most significant predictors were the red band, the SWIR 1 band, CO, PM2.5, and SO_2. Different machine learning techniques

were tested. RFR performed best considering the training dataset, and SVR performed best considering the test dataset. These models allowed us to generate spatial maps identifying areas with a high prevalence of chronic respiratory diseases, representing an effective approach to using RS data in public health research. This work also provides more information about the spatial distribution of respiratory diseases, which can help in the identification and eradication of their possible causes.

7. Overall conclusion and perspectives

The presented PhD project provides an alternative to the use of RS data/techniques in different environmental applications in a region with different environmental and climate conditions. Thus, RS data/techniques help to improve established approaches and can be used to propose new methodologies to retrieve environmental variables and investigate the relationship with health data.

The main objective of this project was to evaluate the applicability of RS data in the study of CRDs, computing the most effective spatial models to estimate and to locate hospital discharge of CRDs between 2013 and 2017 in Quito, Ecuador. The method proposed in this work aimed to generate an empirical spatial LUR model to estimate hospital discharge of CRDs considering dynamic geographic variables. The first step was to evaluate the RS data available in the study area, where most of the images had a high cloud density [32]. Considering this limitation, a new methodology was developed and applied to remove the clouds in order to have more RS data available [152]. After, several spectral indexes were computed.

The second step was to investigate the most adequate RS data to the study area and to this specific problem. NASA EOS satellites were evaluated considering their free data access and availability in the time window of the health data available (2013 to 2017). Specifically, Terra/Aqua MODIS, Landsat-7 ETM+, Landsat-8 OLI were evaluated in order to find the most adequate RS data to predict PM10, and Landsat-8 was selected [6]. Additionally, were concluded that blue and NIR bands are very important as predictors. Several MLTs were also tested (STW, PLS and MLP).

Most of the studies of air pollutants use AOT derived from MODIS products (MOD04-MYD04) [165] as the input in LUR models. This product has a low spatial resolution (3 x 3 km) [166] and in Quito, this resolution is not applicable because the maximum city width is 10 km. Knowing this limitation, new alternatives were investigated using Landsat-8. Some studies have already combined Landsat-8 data with AOT ground stations to model the AOT [142]. However, in our study area this information is not available between 2013 and 2017. Therefore, another alternative is established, considering the visible bands; specifically the blue and red bands; to retrieve AOT [50].

The third step in this research project was to develop different LUR algorithms to retrieve O_3 concentration from RS data, selecting the predictors in order to model air pollutants, also considering Landsat-8 data. A stepwise regression was chosen to select the predictors, based on the comparison with different MLT. The result showed the presence of the coastal aerosol band (B1) and blue band (B2) in the final models, contrasting that

the blue and red bands are related to the AOT presence [10,50]. The SWIR-2 band (B7) is also related to the O_3 concentration in the final model [181,206]. Additionally, other significative predictors are EVI (related to the vegetation) and LST (related to temperature and climate).

In the fourth and final step, based on the previous knowledge, the association between the different CRDs and the environmental parameters computed from RS data were investigated. Spatial CRDs models were developed from different MLT (MLR, MLP, SVR and RFR). The result allowed to map the CRDs presence with RS, air pollution and meteorological variables as predictors. The predictors considered have a known relationship with variables which affect the respiratory health as the vegetation, land use, climate and air pollution [14]. The SVR and RFR were the most effective MLT. It is known that in some classification and regression problems related to RS both techniques are the most efficient [81]. Finally, a relationship between RS data and CRDs were established.

There are some limitations associated with this project, mainly the quantity of satellite images available due to the high cloud density; the quality of air pollutant and meteorological ground data; and the incorrect hospital discharge data.

Considering the limitations of the project, future work still being done. One of the future tasks is to collect more RS data. Another future work will be the improvement of the cloud removal methodologies in order to recover more RS data and then combined different RS data from different satellites. Moreover, in order to have more data available, more years (2018 – 2019) will be used in the establishment of the new models. Another objective will be to use more RS predictors and more multispectral indexes. Finally, the advance of the MLT is real. In this sense, new evaluations to compute more efficient models will be established.

One of the most grateful achievements of this study was the real and established relationship between RS data and the CRDs. Thus, the spatial hospital discharge of CRD maps give a possible answer of the presence of a CRDs. These spatial models can help to local government decision makers to manage the public health and to organize new policies, specifically in places where the highest presence of a CRDs is evident.

The published papers (original files) are presented in Annex I and the conference proceedings (original files) are presented in Annex II.

References

1. WHO Chronic respiratory diseases (CRDs) Available online: https://www.who.int/respiratory/en/ (accessed on Jun 21, 2019).
2. O'Connor, G. T.; Neas, L.; Vaughn, B.; Kattan, M.; Mitchell, H.; Crain, E. F.; Evans, R.; Gruchalla, R.; Morgan, W.; Stout, J.; Adams, G. K.; Lippmann, M. Acute respiratory health effects of air pollution on children with asthma in US inner cities. *J. Allergy Clin. Immunol.* **2008**, *121*, 1133–1139.e1, doi:10.1016/j.jaci.2008.02.020.
3. Wilhelm, M.; Meng, Y.-Y.; Rull, R. P.; English, P.; Balmes, J.; Ritz, B. Environmental Public Health Tracking of Childhood Asthma Using California Health Interview Survey, Traffic, and Outdoor Air Pollution Data. *Environ. Health Perspect.* **2008**, *116*, 1254–1260, doi:10.1289/ehp.10945.
4. Barry, M.; Annesi-Maesano, I. Ten principles for climate, environment and respiratory health. *Eur. Respir. J.* **2017**, *50*, 1701912, doi:10.1183/13993003.01912-2017.
5. Lee, J. H.; Ryu, J. E.; Chung, H. I.; Choi, Y. Y.; Jeon, S. W.; Kim, S. H. Development of spatial scaling technique of forest health sample point information. In *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences - ISPRS Archives*; 2018; Vol. 42, pp. 751–756.
6. Alvarez-Mendoza, C. I.; Teodoro, A.; Torres, N.; Vivanco, V.; Ramirez-Cando, L. Comparison of satellite remote sensing data in the retrieve of PM10 air pollutant over Quito, Ecuador. In *Proceedings of SPIE - The International Society for Optical Engineering*; 2018.
7. NASA NASA's Earth Observing System Available online: https://eospso.gsfc.nasa.gov/ (accessed on Jul 21, 2017).
8. NASA Terra Spacecraft Bus Status. **2014**.
9. Solano, R.; Didan, K.; Jacobson, A.; Huete, A. MODIS Vegetation Index User ' s Guide (MOD13 Series). **2010**, *2010*.
10. Department of the Interior U.S. Geological Survey *Landsat 8 (L8) Data Users Handbook*; 2016; Vol. 8;.
11. Wang, Y.; Du, H.; Xu, Y.; Lu, D.; Wang, X.; Guo, Z. Temporal and spatial variation relationship and influence factors on surface urban heat island and ozone pollution in the Yangtze River Delta, China. *Sci. Total Environ.* **2018**, *631–632*, 921–933, doi:10.1016/J.SCITOTENV.2018.03.050.
12. Herring, J. W. and D. Measuring Vegetation (NDVI & EVI).
13. Chejarla, V. R.; Maheshuni, P. K.; Mandla, V. R. Quantification of LST and CO2 levels using Landsat-8 thermal bands on urban environment. *Geocarto Int.* **2016**, *31*, doi:10.1080/10106049.2015.1094522.

14. Sweileh, W. M.; Al-Jabi, S. W.; Zyoud, S. H.; Sawalha, A. F. Outdoor air pollution and respiratory health: A bibliometric analysis of publications in peer-reviewed journals (1900 - 2017). *Multidiscip. Respir. Med.* **2018**, *13*, 1–12, doi:10.1186/s40248-018-0128-5.

15. US EPA, O. Health Effects of Ozone Pollution Available online: https://www.epa.gov/ozone-pollution/health-effects-ozone-pollution (accessed on May 24, 2018).

16. Kallawicha, K.; Chuang, Y.; Lung, S. C.; Wu, C.; Han, B.; Ting, Y.; Chao, H. J. Outpatient Visits for Allergic Diseases are Associated with Exposure to Ambient Fungal Spores in the Greater Taipei Area. **2018**, 2077–2085, doi:10.4209/aaqr.2018.01.0028.

17. Guarnieri, M.; Balmes, J. R. Outdoor air pollution and asthma. *Lancet (London, England)* **2014**, *383*, 1581–92, doi:10.1016/S0140-6736(14)60617-6.

18. Gupta, P.; Christopher, S. A.; Wang, J.; Gehrig, R.; Lee, Y.; Kumar, N. Satellite remote sensing of particulate matter and air quality assessment over global cities. *Atmos. Environ.* **2006**, *40*, 5880–5892, doi:10.1016/j.atmosenv.2006.03.016.

19. Liu, Y.; Franklin, M.; Kahn, R.; Koutrakis, P. Using aerosol optical thickness to predict ground-level PM2.5 concentrations in the St. Louis area: A comparison between MISR and MODIS. *Remote Sens. Environ.* **2007**, *107*, 33–44, doi:10.1016/j.rse.2006.05.022.

20. Chen, Y.; Han, W.; Chen, S.; Tong, L. Estimating ground-level PM2.5 concentration using Landsat 8 in Chengdu, China. *Proc. SPIE* **2014**, *9259*, 925917–925931, doi:10.1117/12.2068886.

21. Jethva, H.; Torres, O.; Yoshida, Y. Accuracy Assessment of MODIS Land Aerosol Optical Thickness Algorithms using AERONET Measurements. *Atmos. Meas. Tech. Discuss.* **2019**, 1–30, doi:10.5194/amt-2019-77.

22. Li, Z.; Roy, D.; Zhang, H.; Vermote, E.; Huang, H.; Li, Z.; Roy, D. P.; Zhang, H. K.; Vermote, E. F.; Huang, H. Evaluation of Landsat-8 and Sentinel-2A Aerosol Optical Depth Retrievals across Chinese Cities and Implications for Medium Spatial Resolution Urban Aerosol Monitoring. *Remote Sens.* **2019**, *11*, 122, doi:10.3390/rs11020122.

23. Viana, J.; Vasco Santos, J.; Manuel Neiva, R.; Souza, J.; Duarte, L.; Cláudia Teodoro, A.; Freitas, A. Remote Sensing in Human Health: A 10-Year Bibliometric Analysis. *Remote Sens.* **2017**, doi:10.3390/rs9121225.

24. Seltenrich, N. Remote-sensing applications for environmental health research. *Environ. Health Perspect.* **2014**, *122*, A268-75, doi:10.1289/ehp.122-A268.

25. Samson, D. M.; Archer, R. S.; Alimi, T. O.; Arheart, K. L.; Impoinvil, D. E.; Oscar, R.; Fuller, D. O.; Qualls, W. A. New baseline environmental assessment of mosquito ecology in northern Haiti during increased urbanization. *J. Vector Ecol.* **2015**, *40*, 46–58, doi:10.1111/jvec.12131.

26. Ayres-Sampaio, D.; Teodoro, A. C.; Sillero, N.; Santos, C.; Fonseca, J.; Freitas, A. An investigation of the environmental determinants of asthma hospitalizations: An applied spatial approach. *Appl. Geogr.* **2014**, *47*, 10–19, doi:10.1016/j.apgeog.2013.11.011.

27. Alcock, I.; White, M.; Cherrie, M.; Wheeler, B.; Taylor, J.; McInnes, R.; Otte im Kampe, E.; Vardoulakis, S.; Sarran, C.; Soyiri, I.; Fleming, L. Land cover and air pollution are associated with asthma hospitalisations: A cross-sectional study. *Environ. Int.* **2017**, *109*, 29–41, doi:10.1016/J.ENVINT.2017.08.009.

28. Andrusaityte, S.; Grazuleviciene, R.; Kudzyte, J.; Bernotiene, A.; Dedele, A.; Nieuwenhuijsen, M. J. Associations between neighbourhood greenness and asthma in preschool children in Kaunas, Lithuania: a case-control study. *BMJ Open* **2016**, *6*, e010341, doi:10.1136/bmjopen-2015-010341.

29. Fuertes, E.; Standl, M.; Cyrys, J.; Berdel, D.; von Berg, A.; Bauer, C.-P.; Krämer, U.; Sugiri, D.; Lehmann, I.; Koletzko, S.; Carlsten, C.; Brauer, M.; Heinrich, J. A longitudinal analysis of associations between traffic-related air pollution with asthma, allergies and sensitization in the GINIplus and LISAplus birth cohorts. *PeerJ* **2013**, *1*, e193, doi:10.7717/peerj.193.

30. Beelen, R.; Hoek, G.; Vienneau, D.; Eeftens, M.; Dimakopoulou, K.; Pedeli, X.; Tsai, M.-Y.; Künzli, N.; Schikowski, T.; Marcon, A.; Eriksen, K. T.; Raaschou-Nielsen, O.; Stephanou, E.; Patelarou, E.; Lanki, T.; Yli-Tuomi, T.; Declercq, C.; Falq, G.; Stempfelet, M.; Birk, M.; Cyrys, J.; von Klot, S.; Nádor, G.; Varró, M. J.; Dėdelė, A.; Gražulevičienė, R.; Mölter, A.; Lindley, S.; Madsen, C.; Cesaroni, G.; Ranzi, A.; Badaloni, C.; Hoffmann, B.; Nonnemacher, M.; Krämer, U.; Kuhlbusch, T.; Cirach, M.; de Nazelle, A.; Nieuwenhuijsen, M.; Bellander, T.; Korek, M.; Olsson, D.; Strömgren, M.; Dons, E.; Jerrett, M.; Fischer, P.; Wang, M.; Brunekreef, B.; de Hoogh, K. Development of NO2 and NOx land use regression models for estimating air pollution exposure in 36 study areas in Europe – The ESCAPE project. *Atmos. Environ.* **2013**, *72*, 10–23, doi:10.1016/J.ATMOSENV.2013.02.037.

31. Cilluffo, G.; Ferrante, G.; Fasola, S.; Montalbano, L.; Malizia, V.; Piscini, A.; Romaniello, V.; Silvestri, M.; Stramondo, S.; Stafoggia, M.; Ranzi, A.; Viegi, G.; La Grutta, S. Associations of greenness, greyness and air pollution exposure with children's health: a cross-sectional study in Southern Italy. *Environ. Heal.* **2018**, *17*, 86, doi:10.1186/s12940-018-0430-x.

32. Alvarez, C. I.; Teodoro, A.; Tierra, A. Evaluation of automatic cloud removal method for high elevation areas in Landsat 8 OLI images to improve environmental indexes computation. In *Proc.SPIE*; 2017; Vol. 10428.

33. Alvarez-Mendoza, C. I.; Teodoro, A.; Ramirez-Cando, L. Improving NDVI by removing cirrus clouds with optical remote sensing data from Landsat-8 – A case study in Quito, Ecuador. *Remote*

Sens. Appl. Soc. Environ. **2019**, doi:10.1016/j.rsase.2018.11.008.

34. Alvarez-Mendoza, C. I.; Teodoro, A.; Ramirez-Cando, L. Spatial estimation of surface ozone concentrations in Quito Ecuador with remote sensing data, air pollution measurements and meteorological variables. *Environ. Monit. Assess.* **2019**, *191*, 155, doi:10.1007/s10661-019-7286-6.
35. Xu, M.; Jia, X.; Pickering, M. Automatic cloud removal for Landsat 8 OLI images using cirrus band. *Int. Geosci. Remote Sens. Symp.* **2014**, 2511–2514, doi:10.1109/IGARSS.2014.6946983.
36. Alvarez-Mendoza, C. I.; Teodoro, A. C.; Torres, N.; Vivanco, V.; Alvarez-Mendoza, C. I.; Teodoro, A. C.; Torres, N.; Vivanco, V. Assessment of Remote Sensing Data to Model PM10 Estimation in Cities with a Low Number of Air Quality Stations: A Case of Study in Quito, Ecuador. *Environ. 2019, Vol. 6, Page 85* **2019**, *6*, 85, doi:10.3390/ENVIRONMENTS6070085.
37. Baldock, J. W. *Geology of Ecuador: Explanatory Bulletin of the National Geological Map of the Republic of Ecuador : 1:1,000,000 Scale*; Ministerio de Recursos Naturales y Energéticos, Dirección General de Geolog{\'\i}a y Minas, 1982;
38. Secretaria del Ambiente de Quito Red Metropolitana de Monitoreo Atmosférico de Quito Available online: http://www.quitoambiente.gob.ec/ambiente/index.php/generalidades (accessed on Jun 26, 2018).
39. Rees, W. G. *Physical Principles of Remote Sensing*; Cambridge University Press, 2013; ISBN 9781139851374.
40. Chuvieco, E. *Fundamentals of Satellite Remote Sensing: An Environmental Approach, Second Edition*; CRC Press, 2016; ISBN 9781498728072.
41. U.S. Geological Survey *Landsat—Earth observation satellites*; Version 1.; Reston, VA, 2015; Vol. 2015–3081;.
42. NASA, U. *Preliminary Assessment of the Value of Landsat 7 ETM+ Data following Scan Line Corrector Malfunction Executive Summary*; 2003;
43. Girard, C. M.; Girard, M. C. *Processing of Remote Sensing Data*; Taylor & Francis, 2003; ISBN 9789058092328.
44. Sahu, K. C. *Textbook of Remote Sensing and Geographical Information Systems*; Atlantic Publishers & Distributors (P) Limited, 2007; ISBN 9788126909094.
45. Chavez, P. S. An improved dark-object subtraction technique for atmospheric scattering correction of multispectral data. *Remote Sens. Environ.* **1988**, *24*, 459–479, doi:10.1016/0034-4257(88)90019-3.
46. Vet-Mote, E. F.; Tan&, D.; Deuzc, J. L.; Herman, M.; Morcrette, J.-J. Second Simulation of the

Satellite Signal in the Solar Spectrum, 6s: An Overview. *IEEE Trans. Geosci. Remote Sens.* **1997**, *35*.

47. Cooley, T.; Anderson, G. P.; Felde, G. W.; Hoke, M. L.; Ratkowski, A. J.; Chetwynd, J. H.; Gardner, J. A.; Adler-Golden, S. M.; Matthew, M. W.; Berk, A.; Bernstein, L. S.; Acharya, P. K.; Miller, D.; Lewis, P. FLAASH, a MODTRAN4-based atmospheric correction algorithm, its application and validation. In *IEEE International Geoscience and Remote Sensing Symposium*; IEEE; Vol. 3, pp. 1414–1418.
48. Allred, C. L.; Jeong, L. S.; Chetwynd, J. H. Flaash, a modtran4 atmospheric correction package. **1994**, *4*.
49. Mandanici, E.; Franci, F.; Bitelli, G.; Agapiou, A.; Alexakis, D.; Hadjimitsis, D. G. Comparison between empirical and physically based models of atmospheric correction. *Proc. SPIE - Int. Soc. Opt. Eng.* **2015**, *9535*, 95350E, doi:10.1117/12.2193176.
50. Vermote, E.; Justice, C.; Claverie, M.; Franch, B. Preliminary analysis of the performance of the Landsat 8/OLI land surface reflectance product. *Remote Sens. Environ.* **2016**, *185*, 46–56, doi:10.1016/J.RSE.2016.04.008.
51. USGS *Landsat 8 Surface Reflectance Code (LaSCR) Product Guide*; South Dakota, 2019;
52. Roger, P. J. C.; Vermote, E. F.; Ray, J. P. *MODIS Surface Reflectance User ' s Guide*; 2015;
53. Xue, J.; Su, B. Significant Remote Sensing Vegetation Indices: A Review of Developments and Applications. *J. Sensors* **2017**, *2017*, 1–17, doi:10.1155/2017/1353691.
54. Tucker, C. J. Red and photographic infrared linear combinations for monitoring vegetation. *Remote Sens. Environ.* **1979**, *8*, 127–150, doi:10.1016/0034-4257(79)90013-0.
55. Huete, A. . A soil-adjusted vegetation index (SAVI). *Remote Sens. Environ.* **1988**, *25*, 295–309, doi:10.1016/0034-4257(88)90106-X.
56. Huete, A.; Didan, K.; Miura, T.; Rodriguez, E. .; Gao, X.; Ferreira, L. . Overview of the radiometric and biophysical performance of the MODIS vegetation indices. *Remote Sens. Environ.* **2002**, *83*, 195–213, doi:10.1016/S0034-4257(02)00096-2.
57. As-syakur, A. R.; Adnyana, I. W. S.; Arthana, I. W.; Nuarsa, I. W. Enhanced Built-Up and Bareness Index (EBBI) for Mapping Built-Up and Bare Land in an Urban Area. *Remote Sens.* **2012**, *4*, 2957–2970, doi:10.3390/rs4102957.
58. Li, Z.-L.; Duan, S.-B. Land Surface Temperature. *Compr. Remote Sens.* **2018**, 264–283, doi:10.1016/B978-0-12-409548-9.10375-6.
59. Vieira, D.; Teodoro, A.; Gomes, A. Analysing Land Surface Temperature variations during Fogo Island (Cape Verde) 2014-2015 eruption with Landsat 8 images. *Proc. SPIE* **2016**, *10005*,

1000508, doi:10.1117/12.2241382.

60. Gillespie, A. Land surface emissivity. In *Encyclopedia of Remote Sensing*; Njoku, E., Ed.; Springer-Verlag New York, 2014; p. 939 ISBN 978-0-387-36698-2.
61. Estallo, E. L.; Benitez, E. M.; Lanfri, M. A.; Scavuzzo, C. M.; Almirón, W. R. MODIS Environmental Data to Assess Chikungunya, Dengue, and Zika Diseases Through Aedes (Stegomia) aegypti Oviposition Activity Estimation. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2016**, *9*, 5461–5466, doi:10.1109/JSTARS.2016.2604577.
62. Liang, C.-S.; Liu, H.; He, K.-B.; Ma, Y.-L. Assessment of regional air quality by a concentration-dependent Pollution Permeation Index OPEN. **2016**, doi:10.1038/srep34891.
63. Al-Hamdan, A. Z.; Albashaireh, R. N.; Al-Hamdan, M. Z.; Crosson, W. L. The association of remotely sensed outdoor fine particulate matter with cancer incidence of respiratory system in the USA. *J. Environ. Sci. Heal. Part A* **2017**, *52*, 547–554, doi:10.1080/10934529.2017.1284432.
64. Ai, S.; Qian, Z. M.; Guo, Y.; Yang, Y.; Rolling, C. A.; Liu, E.; Wu, F.; Lin, H. Long-term exposure to ambient fine particles associated with asthma: A cross-sectional study among older adults in six low- and middle-income countries. *Environ. Res.* **2019**, *168*, 141–145, doi:10.1016/j.envres.2018.09.028.
65. Li, L.; Hart, J. E.; Coull, B. A.; Cao, S.-J.; Spengler, J. D.; Adamkiewicz, G. Effect of Residential Greenness and Nearby Parks on Respiratory and Allergic Diseases among Middle School Adolescents in a Chinese City. *Int. J. Environ. Res. Public Health* **2019**, *16*, doi:10.3390/ijerph16060991.
66. Ielpo, P.; Paolillo, V.; de Gennaro, G.; Dambruoso, P. R. PM10 and gaseous pollutants trends from air quality monitoring networks in Bari province: principal component analysis and absolute principal component scores on a two years and half data set. *Chem. Cent. J.* **2014**, *8*, 14, doi:10.1186/1752-153X-8-14.
67. Pope, R.; Wu, J. A multi-objective assessment of an air quality monitoring network using environmental, economic, and social indicators and GIS-based models. *J. Air Waste Manage. Assoc.* **2014**, *64*, 721–737, doi:10.1080/10962247.2014.888378.
68. WHO ICD-10 Version:2016 Available online: https://icd.who.int/browse10/2016/en (accessed on Jun 23, 2019).
69. Gilliland, F.; Avol, E.; Kinney, P.; Jerrett, M.; Dvonch, T.; Lurmann, F.; Buckley, T.; Breysse, P.; Keeler, G.; de Villiers, T.; McConnell, R. Air Pollution Exposure Assessment for Epidemiologic Studies of Pregnant Women and Children: Lessons Learned from the Centers for Children's Environmental Health and Disease Prevention Research. *Environ. Health Perspect.* **2005**, *113*,

1447–1454, doi:10.1289/ehp.7673.

70. Ryan, P. H.; LeMasters, G. K. A review of land-use regression models for characterizing intraurban air pollution exposure. *Inhal. Toxicol.* **2007**, doi:10.1080/08958370701495998.
71. Alpaydin, E. *Introduction to Machine Learning*; Adaptive Computation and Machine Learning series; MIT Press, 2014; ISBN 9780262028189.
72. MathWorks What Is Machine Learning? | How It Works Available online: https://www.mathworks.com/discovery/machine-learning.html (accessed on Jul 14, 2019).
73. Yamashita, T.; Yamashita, K.; Kamimura, R. A Stepwise AIC Method for Variable Selection in Linear Regression. *Commun. Stat. - Theory Methods* **2007**, *36*, 2395–2403, doi:10.1080/03610920701215639.
74. NCSS; LLC *Stepwise Regression*;
75. Leach, A. R.; Gillet, V. J. *An Introduction to Chemoinformatics*; Springer ebook collection / Chemistry and Materials Science 2005-2008; Springer, 2007; ISBN 9781402062902.
76. Williams, L. J.; Abdi, H.; Williams, L. J. Partial Least Squares Methods: Partial Least Squares Correlation and Partial Least Square Regression. In *Computational Toxicology: Volume II*; Reisfeld, B., Mayeno, A. N., Eds.; Humana Press: Totowa, NJ, 2013; Vol. 930, pp. 549–579 ISBN 978-1-62703-059-5.
77. Mather, P.; Tso, B. *Classification Methods for Remotely Sensed Data*; CRC Press, 2003; ISBN 9780203303566.
78. Tzanakou, E. M. *Supervised and Unsupervised Pattern Recognition: Feature Extraction and Computational Intelligence*; Industrial Electronics; CRC Press, 2017; ISBN 9781420049770.
79. Murtagh, F. Multilayer perceptrons for classification and regression. *Neurocomputing* **1991**, *2*, 183–197, doi:10.1016/0925-2312(91)90023-5.
80. Vapnik, V. *The Nature of Statistical Learning Theory*; Information Science and Statistics; Springer New York, 2013; ISBN 9781475732641.
81. Yang, F.; Ichii, K.; White, M. A.; Hashimoto, H.; Michaelis, A. R.; Votava, P.; Zhu, A.-X.; Huete, A.; Running, S. W.; Nemani, R. R. Developing a continental-scale measure of gross primary production by combining MODIS and AmeriFlux data through Support Vector Machine approach. *Remote Sens. Environ.* **2007**, *110*, 109–122, doi:10.1016/J.RSE.2007.02.016.
82. Vapnik, V. N. *Statistical learning theory*; Adaptive and learning systems for signal processing, communications, and control; Wiley, 1998; ISBN 9780471030034.
83. Ho, T. K. A Theory of Multiple Classifier Systems And Its Application to Visual Word Recognition 1992.

84. Chen, C. H. *Signal and Image Processing for Remote Sensing*; Signal and Image Processing for Remote Sensing; Taylor & Francis, 2006; ISBN 9780849350917.

85. Ghorbanzadeh, O.; Blaschke, T.; Gholamnia, K.; Meena, S.; Tiede, D.; Aryal, J.; Ghorbanzadeh, O.; Blaschke, T.; Gholamnia, K.; Meena, S. R.; Tiede, D.; Aryal, J. Evaluation of Different Machine Learning Methods and Deep-Learning Convolutional Neural Networks for Landslide Detection. *Remote Sens.* **2019**, *11*, 196, doi:10.3390/rs11020196.

86. Rodriguez-Galiano, V.; Sanchez-Castillo, M.; Chica-Olmo, M.; Chica-Rivas, M. Machine learning predictive models for mineral prospectivity: An evaluation of neural networks, random forest, regression trees and support vector machines. *Ore Geol. Rev.* **2015**, *71*, 804–818, doi:10.1016/J.OREGEOREV.2015.01.001.

87. Rodriguez-Galiano, V. F.; Sanchez-Castillo, M.; Dash, J.; Atkinson, P. M.; Ojeda-Zujar, J. Modelling interannual variation in the spring and autumn land surface phenology of the European forest. *Biogeosciences* **2016**, *13*, 3305–3317, doi:10.5194/bg-13-3305-2016.

88. Dr. S. C. Liew Principles of Remote Sensing Available online: http://www.crisp.nus.edu.sg/~research/tutorial/optical.htm (accessed on Jul 21, 2017).

89. Ju, J.; Roy, D. P. The availability of cloud-free Landsat ETM+ data over the conterminous United States and globally. *Remote Sens. Environ.* **2008**, *112*, 1196–1211, doi:10.1016/j.rse.2007.08.011.

90. Asner, G. P. Cloud cover in Landsat observations of the Brazilian Amazon. *Int. J. Remote Sens.* **2001**, *22*, 3855–3862, doi:10.1080/01431160010006926.

91. Fernández, G.; Obermeier, W.; Gerique, A.; Sandoval, M.; Lehnert, L.; Thies, B.; Bendix, J. Land Cover Change in the Andes of Southern Ecuador—Patterns and Drivers. *Remote Sens.* **2015**, *7*, 2509–2542, doi:10.3390/rs70302509.

92. Herring, J. W. and D. Measuring Vegetation (NDVI & EVI) Available online: https://earthobservatory.nasa.gov/Features/MeasuringVegetation/measuring_vegetation_1.php (accessed on Jul 12, 2017).

93. Rajitha, K.; Prakash Mohan, M. M.; Varma, M. R. R. Effect of cirrus cloud on normalized difference Vegetation Index (NDVI) and Aerosol Free Vegetation Index (AFRI): A study based on LANDSAT 8 images. *ICAPR 2015 - 2015 8th Int. Conf. Adv. Pattern Recognit.* **2015**, 2–6, doi:10.1109/ICAPR.2015.7050710.

94. Richter, R.; Wang, X.; Bachmann, M.; Schläpfer, D. Correction of cirrus effects in Sentinel-2 type of imagery. *Int. J. Remote Sens.* **2011**, *32*, 2931–2941, doi:10.1080/01431161.2010.520346.

95. Shen, H.; Li, H.; Qian, Y.; Zhang, L.; Yuan, Q. An effective thin cloud removal procedure for visible remote sensing images. *ISPRS J. Photogramm. Remote Sens.* **2014**, *96*, 224–235,

doi:10.1016/J.ISPRSJPRS.2014.06.011.

96. Gao, B.-C.; Li, R.-R.; Gao, B.-C.; Li, R.-R. Removal of Thin Cirrus Scattering Effects in Landsat 8 OLI Images Using the Cirrus Detecting Channel. *Remote Sens.* **2017**, *9*, 834, doi:10.3390/rs9080834.

97. Lv, H.; Wang, Y.; Yang, Y. Modeling of Thin-Cloud TOA Reflectance Using Empirical Relationships and Two Landsat-8 Visible Band Data. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 839–850, doi:10.1109/TGRS.2018.2861939.

98. USGS User Guide Landsat 8 Operational Land Imager (Oli). **2013**, 1–16.

99. Lv, H.; Wang, Y.; Shen, Y. An empirical and radiative transfer model based algorithm to remove thin clouds in visible bands. *Remote Sens. Environ.* **2016**, *179*, 183–195, doi:10.1016/j.rse.2016.03.034.

100. Cheng, Q.; Shen, H.; Zhang, L.; Yuan, Q.; Zeng, C. Cloud removal for remotely sensed images by similar pixel replacement guided with a spatio-temporal MRF model. *ISPRS J. Photogramm. Remote Sens.* **2014**, *92*, 54–68, doi:10.1016/j.isprsjprs.2014.02.015.

101. Bo-Cai Gao; Rong-Rong Li Removal of Thin Cirrus Scattering Effects for Remote Sensing of Ocean Color From Space. *IEEE Geosci. Remote Sens. Lett.* **2012**, *9*, 972–976, doi:10.1109/LGRS.2012.2187876.

102. Shen, Y.; Wang, Y.; Lv, H.; Qian, J. Removal of thin clouds in landsat-8 OLI data with independent component analysis. *Remote Sens.* **2015**, *7*, 11481–11500, doi:10.3390/rs70911481.

103. Huadong, D.; Yongqi, W.; Yaming, C. Studies on Cloud Detection of Atmospheric Remote Sensing Image Using ICA Algorithm. *Computer (Long. Beach. Calif).* **2009**, 1–4.

104. Zhu, Z.; Woodcock, C. E. Object-based cloud and cloud shadow detection in Landsat imagery. *Remote Sens. Environ.* **2012**, *118*, 83–94, doi:10.1016/j.rse.2011.10.028.

105. Stephens, G. L. Cloud Feedbacks in the Climate System: A Critical Review. *J. Clim.* **2005**, *18*, 237–273, doi:10.1175/JCLI-3243.1.

106. Gao, B.; Kaufman, Y. J.; Han, W.; Wiscombe, W. J. Spectral Region Using the Sensitive 1 . 375 Cirrus Detecting Channel. **1998**, *103*, 169–176.

107. Lin, C.-H.; Lai, K.-H.; Chen, Z.-B.; Chen, J.-Y. Patch-Based Information Reconstruction of Cloud-Contaminated Multitemporal Images. *IEEE Trans. Geosci. Remote Sens.* **2014**, *52*, 163–174, doi:10.1109/TGRS.2012.2237408.

108. Wu, M.; Wu, C.; Huang, W.; Niu, Z.; Wang, C.; Li, W.; Hao, P. An improved high spatial and temporal data fusion approach for combining Landsat and MODIS data to generate daily synthetic Landsat imagery. *Inf. Fusion* **2016**, *31*, 14–25, doi:10.1016/j.inffus.2015.12.005.

109. Shen, Y.; Wang, Y.; Lv, H.; Li, H. Removal of Thin Clouds Using Cirrus and QA Bands of Landsat-8. *Photogramm. Eng. Remote Sens.* **2015**, *81*, 721–731, doi:10.14358/PERS.81.9.721.

110. Hyvärinen, A.; Oja, E. Independent component analysis: algorithms and applications. *Neural networks* **2000**, *13*, 411–430, doi:10.1016/S0893-6080(00)00026-5.

111. Instituto Nacional de Meteorología e Hidrología *Boletín Climatológico Anual 2015*; 2016;

112. R Core Team R: A Language and Environment for Statistical Computing 2016.

113. Hijmans, R. J. raster: Geographic Data Analysis and Modeling 2016.

114. Bivand, R.; Keitt, T.; Rowlingson, B. rgdal: Bindings for the Geospatial Data Abstraction Library 2016.

115. Greenberg, J. A.; Mattiuzzi, M. gdalUtils: Wrappers for the Geospatial Data Abstraction Library (GDAL) Utilities 2015.

116. Ji, C. Y. Haze reduction from the visible bands of LANDSAT TM and ETM+ images over a shallow water reef environment. *Remote Sens. Environ.* **2008**, *112*, 1773–1783, doi:10.1016/j.rse.2007.09.006.

117. Hyvärinen, A.; Karhunen, J.; Oja, E. Independent Component Analysis. *Appl. Comput. Harmon. Anal.* **2001**, *21*, 135–144, doi:10.1002/0471221317.

118. ENVI ENVI Atmospheric Correction Module: QUAC and FLAASH user's guide. *Modul. Version* **2009**, 44.

119. Roy, D. P.; Kovalskyy, V.; Zhang, H. K.; Vermote, E. F.; Yan, L.; Kumar, S. S.; Egorov, A. Characterization of Landsat-7 to Landsat-8 reflective wavelength and normalized difference vegetation index continuity. *Remote Sens. Environ.* **2016**, *185*, 57–70, doi:10.1016/j.rse.2015.12.024.

120. Mishra, N. B.; Mainali, K. P. Greening and browning of the Himalaya: Spatial patterns and the role of climatic change and human drivers. *Sci. Total Environ.* **2017**, *587–588*, 326–339, doi:10.1016/j.scitotenv.2017.02.156.

121. Kuenzer, C.; Dech, S.; Wagner, W. Remote Sensing Time Series Revealing Land Surface Dynamics: Status Quo and the Pathway Ahead. In; Springer, Cham, 2015; pp. 1–24.

122. Zambrano, F.; Lillo-Saavedra, M.; Verbist, K.; Lagos, O. Sixteen Years of Agricultural Drought Assessment of the BioBío Region in Chile Using a 250 m Resolution Vegetation Condition Index (VCI). *Remote Sens.* **2016**, *8*, 530, doi:10.3390/rs8060530.

123. WHO Ambient (outdoor) air quality and health Available online: http://www.who.int/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health (accessed on Aug 30, 2018).

124. Kutlar Joss, M.; Eeftens, M.; Gintowt, E.; Kappeler, R.; Künzli, N. Time to harmonize national

ambient air quality standards. *Int. J. Public Health* **2017**, *62*, 453–462, doi:10.1007/s00038-017-0952-y.

125. Kobza, J.; Geremek, M.; Dul, L. Characteristics of air quality and sources affecting high levels of PM10 and PM2.5 in Poland, Upper Silesia urban area. *Environ. Monit. Assess.* **2018**, *190*, 515, doi:10.1007/s10661-018-6797-x.
126. Health Organization, W.; Office for Europe, R. *Health effects of particulate matter*; 2013;
127. Capezzuto, L.; Abbamonte, L.; De Vito, S.; Massera, E.; Formisano, F.; Fattoruso, G.; Di Francia, G.; Buonanno, A. A maker friendly mobile and social sensing approach to urban air quality monitoring. In *IEEE SENSORS 2014 Proceedings*; IEEE, 2014; pp. 12–16.
128. Hasenfratz, D.; Saukh, O.; Walser, C.; Hueglin, C.; Fierz, M.; Thiele, L. Pushing the spatio-temporal resolution limit of urban air pollution maps. In *2014 IEEE International Conference on Pervasive Computing and Communications (PerCom)*; IEEE, 2014; pp. 69–77.
129. Alvarez, C. I.; Padilla Almeida, O.; Álvarez Mendoza, C. I.; Padilla Almeida, O. Estimación de la contaminación del aire por PM10 en Quito a través de índices ambientales con imágenes LANDSAT ETM+. *Rev. Cart.* **2016**, 135–147.
130. Cevallos, V. M.; Díaz, V.; Sirois, C. M. Particulate matter air pollution from the city of Quito, Ecuador, activates inflammatory signaling pathways in vitro. *Innate Immun.* **2017**, *23*, 392–400, doi:10.1177/1753425917699864.
131. Raysoni, A. U.; Armijos, R. X.; Weigel, M. M.; Montoya, T.; Eschanique, P.; Racines, M.; Li, W.-W. Assessment of indoor and outdoor PM species at schools and residences in a high-altitude Ecuadorian urban center. *Environ. Pollut.* **2016**, *214*, 668–679, doi:10.1016/J.ENVPOL.2016.04.085.
132. Yang, X.; Zheng, Y.; Geng, G.; Liu, H.; Man, H.; Lv, Z.; He, K.; de Hoogh, K. Development of PM2.5 and NO2 models in a LUR framework incorporating satellite remote sensing and air quality model data in Pearl River Delta region, China. *Environ. Pollut.* **2017**, *226*, 143–153, doi:10.1016/J.ENVPOL.2017.03.079.
133. Stafoggia, M.; Schwartz, J.; Badaloni, C.; Bellander, T.; Alessandrini, E.; Cattani, G.; de' Donato, F.; Gaeta, A.; Leone, G.; Lyapustin, A.; Sorek-Hamer, M.; de Hoogh, K.; Di, Q.; Forastiere, F.; Kloog, I. Estimation of daily PM10 concentrations in Italy (2006–2012) using finely resolved satellite data, land use variables and meteorology. *Environ. Int.* **2017**, *99*, 234–244, doi:10.1016/J.ENVINT.2016.11.024.
134. Shi, Y.; Lau, K. K.-L.; Ng, E. Incorporating wind availability into land use regression modelling of air quality in mountainous high-density urban environment. *Environ. Res.* **2017**, *157*, 17–29,

doi:10.1016/J.ENVRES.2017.05.007.

135. Son, Y.; Osornio-Vargas, Á. R.; O'Neill, M. S.; Hystad, P.; Texcalac-Sangrador, J. L.; Ohman-Strickland, P.; Meng, Q.; Schwander, S. Land use regression models to assess air pollution exposure in Mexico City using finer spatial and temporal input parameters. *Sci. Total Environ.* **2018**, *639*, 40–48, doi:10.1016/J.SCITOTENV.2018.05.144.

136. Zou, B.; Chen, J.; Zhai, L.; Fang, X.; Zheng, Z.; Zou, B.; Chen, J.; Zhai, L.; Fang, X.; Zheng, Z. Satellite Based Mapping of Ground PM2.5 Concentration Using Generalized Additive Modeling. *Remote Sens.* **2016**, *9*, 1, doi:10.3390/rs9010001.

137. Wu, C.-D.; Chen, Y.-C.; Pan, W.-C.; Zeng, Y.-T.; Chen, M.-J.; Guo, Y. L.; Lung, S.-C. C. Land-use regression with long-term satellite-based greenness index and culture-specific sources to model PM2.5 spatial-temporal variability. *Environ. Pollut.* **2017**, *224*, 148–157, doi:10.1016/J.ENVPOL.2017.01.074.

138. He, J.; Zha, Y.; Zhang, J.; Gao, J. Aerosol indices derived from MODIS data for indicating aerosol-induced air pollution. *Remote Sens.* **2014**, *6*, 1587–1604, doi:10.3390/rs6021587.

139. Just, A.; De Carli, M.; Shtein, A.; Dorman, M.; Lyapustin, A.; Kloog, I.; Just, A. C.; De Carli, M. M.; Shtein, A.; Dorman, M.; Lyapustin, A.; Kloog, I. Correcting Measurement Error in Satellite Aerosol Optical Depth with Machine Learning for Modeling PM2.5 in the Northeastern USA. *Remote Sens.* **2018**, *10*, 803, doi:10.3390/rs10050803.

140. Wan, Z.; others MODIS land surface temperature products users guide. *Inst. Comput. Earth Syst. Sci. Univ. Calif. St. Barbar. CA, USA* **2013**.

141. Olmanson, L. G.; Brezonik, P. L.; Finlay, J. C.; Bauer, M. E. Comparison of Landsat 8 and Landsat 7 for regional measurements of CDOM and water clarity in lakes. *Remote Sens. Environ.* **2016**, *185*, 119–128, doi:10.1016/j.rse.2016.01.007.

142. Bilal, M.; Nichol, J. E.; Bleiweiss, M. P.; Dubois, D. A Simplified high resolution MODIS aerosol retrieval algorithm (SARA) for use over mixed surfaces. *Remote Sens. Environ.* **2013**, *136*, 135–145, doi:10.1016/j.rse.2013.04.014.

143. Meng, X.; Fu, Q.; Ma, Z.; Chen, L.; Zou, B.; Zhang, Y.; Xue, W.; Wang, J.; Wang, D.; Kan, H.; Liu, Y. Estimating ground-level PM10 in a Chinese city by combining satellite data, meteorological information and a land use regression model. *Environ. Pollut.* **2016**, *208*, 177–184, doi:10.1016/J.ENVPOL.2015.09.042.

144. Shahraiyni, H. T.; Sodoudi, S. Statistical modeling approaches for pm10prediction in urban areas; A review of 21st-century studies. *Atmosphere (Basel).* **2016**, *7*, 10–13, doi:10.3390/atmos7020015.

145. Naughton, O.; Donnelly, A.; Nolan, P.; Pilla, F.; Misstear, B. D.; Broderick, B. A land use regression model for explaining spatial variation in air pollution levels using a wind sector based approach. *Sci. Total Environ.* **2018**, *630*, 1324–1334, doi:10.1016/J.SCITOTENV.2018.02.317.

146. Li, X.; Zhang, Y.; Bao, Y.; Luo, J.; Jin, X.; Xu, X.; Song, X.; Yang, G. Exploring the Best Hyperspectral Features for LAI Estimation Using Partial Least Squares Regression. *Remote Sens.* **2014**, *6*, 6221–6241, doi:10.3390/rs6076221.

147. Chen, G.; Meentemeyer, R. Remote Sensing of Forest Damage by Diseases and Insects. In *Remote Sensing for Sustainability*; Weng, Q., Ed.; Remote Sensing Applications Series; CRC Press: Boca Raton, Florida, 2016; p. 357 ISBN 9781315354644.

148. Xu, W.; Riley, E. A.; Austin, E.; Sasakura, M.; Schaal, L.; Gould, T. R.; Hartin, K.; Simpson, C. D.; Sampson, P. D.; Yost, M. G.; Larson, T. V.; Xiu, G.; Vedal, S. Use of mobile and passive badge air monitoring data for NO X and ozone air pollution spatial exposure prediction models. *J. Expo. Sci. Environ. Epidemiol.* **2017**, *27*, 184–192, doi:10.1038/jes.2016.9.

149. Rosero-Vlasova, O. A.; Vlassova, L.; Pérez-Cabello, F.; Montorio, R.; Nadal-Romero, E. Modeling soil organic matter and texture from satellite data in areas affected by wildfires and cropland abandonment in Aragón, Northern Spain. *J. Appl. Remote Sens.* **2018**, *12*, 1, doi:10.1117/1.JRS.12.042803.

150. Liu, W.; Li, X.; Chen, Z.; Zeng, G.; León, T.; Liang, J.; Huang, G.; Gao, Z.; Jiao, S.; He, X.; Lai, M. Land use regression models coupled with meteorology to model spatial and temporal variability of NO2 and PM10 in Changsha, China. *Atmos. Environ.* **2015**, *116*, 272–280, doi:10.1016/J.ATMOSENV.2015.06.056.

151. Gardner, M. .; Dorling, S. . Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. *Atmos. Environ.* **1998**, *32*, 2627–2636, doi:10.1016/S1352-2310(97)00447-0.

152. Alvarez-Mendoza, C. I.; Teodoro, A.; Ramirez-Cando, L. Improving NDVI by removing cirrus clouds with optical remote sensing data from Landsat-8 – A case study in Quito, Ecuador. *Remote Sens. Appl. Soc. Environ.* **2018**, *13*, 257–274, doi:10.1016/j.rsase.2018.11.008.

153. Othman, N.; Jafri, M. Z. M.; San, L. H. Estimating particulate matter concentration over arid region using satellite remote sensing: A case study in Makkah, Saudi Arabia. *Mod. Appl. Sci.* **2010**, *4*, 131, doi:10.5539/mas.v4n11p131.

154. Bilguunmaa, M.; Batbayar, J.; Tuya, S. Estimation of PM10 concentration using satellite data in Ulaanbaatar City. *SPIE Asia Pacific Remote Sens.* **2014**, 92591O--92591O, doi:10.1117/12.2069149.

155. Ángel, M.; Gutiérrez, R. Uso de Modelos Lineales Generalizados (MLG) para la interpolación espacial de PM10 utilizando imágenes satelitales Landsat para la ciudad de Bogotá , Colombia. **2017**, *22*, 105–121, doi:10.19053/01233769.5600.

156. Ghaleb, F.; Mario, M.; Sandra, A. Regional Landsat-Based Drought Monitoring from 1982 to 2014. *Climate* **2015**, *3*, 563–577, doi:10.3390/cli3030563.

157. Sobrino, J. A.; Jiménez-Muñoz, J. C.; Sòria, G.; Romaguera, M.; Guanter, L.; Moreno, J.; Plaza, A.; Martínez, P. Land surface emissivity retrieval from different VNIR and TIR sensors. *IEEE Trans. Geosci. Remote Sens.* **2008**, *46*, 316–327, doi:10.1109/TGRS.2007.904834.

158. Li, S.; Jiang, G. M. Land Surface Temperature Retrieval from Landsat-8 Data with the Generalized Split-Window Algorithm. *IEEE Access* **2018**, *6*, 18149–18162, doi:10.1109/ACCESS.2018.2818741.

159. Habermann, M.; Billger, M.; Haeger-Eugensson, M. Land use Regression as Method to Model Air Pollution. Previous Results for Gothenburg/Sweden. *Procedia Eng.* **2015**, *115*, 21–28, doi:10.1016/J.PROENG.2015.07.350.

160. Zhang, J. J. Y.; Sun, L.; Barrett, O.; Bertazzon, S.; Underwood, F. E.; Johnson, M. Development of land-use regression models for metals associated with airborne particulate matter in a North American city. *Atmos. Environ.* **2015**, *106*, 165–177, doi:10.1016/J.ATMOSENV.2015.01.008.

161. Sheela, K. G.; Deepa, S. N. Review on Methods to Fix Number of Hidden Neurons in Neural Networks. *Math. Probl. Eng.* **2013**, *2013*, 1–11, doi:10.1155/2013/425740.

162. Cattani, G.; Gaeta, A.; Di Menno di Bucchianico, A.; De Santis, A.; Gaddi, R.; Cusano, M.; Ancona, C.; Badaloni, C.; Forastiere, F.; Gariazzo, C.; Sozzi, R.; Inglessis, M.; Silibello, C.; Salvatori, E.; Manes, F.; Cesaroni, G. Development of land-use regression models for exposure assessment to ultrafine particles in Rome, Italy. *Atmos. Environ.* **2017**, *156*, 52–60, doi:10.1016/J.ATMOSENV.2017.02.028.

163. Wang, M.; Sampson, P. D.; Hu, J.; Kleeman, M.; Keller, J. P.; Olives, C.; Szpiro, A. A.; Vedal, S.; Kaufman, J. D. Combining Land-Use Regression and Chemical Transport Modeling in a Spatiotemporal Geostatistical Model for Ozone and PM 2.5. *Environ. Sci. Technol.* **2016**, *50*, 5111–5118, doi:10.1021/acs.est.5b06001.

164. Alexeeff, S. E.; Schwartz, J.; Kloog, I.; Chudnovsky, A.; Koutrakis, P.; Coull, B. A. Consequences of kriging and land use regression for PM2.5 predictions in epidemiologic analyses: Insights into spatial variability using high-resolution satellite data. *J. Expo. Sci. Environ. Epidemiol.* **2015**, *25*, 138–144, doi:10.1038/jes.2014.40.

165. Beloconi, A.; Chrysoulakis, N.; Lyapustin, A.; Utzinger, J.; Vounatsou, P. Bayesian geostatistical

modelling of PM10 and PM2.5 surface level concentrations in Europe using high-resolution satellite-derived products. *Environ. Int.* **2018**, *121*, 57–70, doi:10.1016/j.envint.2018.08.041.

166. Remer, L. A.; Mattoo, S.; Levy, R. C.; Munchak, L. A. MODIS 3 km aerosol product: algorithm and global perspective. *Atmos. Meas. Tech.* **2013**, *6*, 1829–1844, doi:10.5194/amt-6-1829-2013.
167. Teodoro, A. A study on the Quality of the Vegetation Index obtainded from MODIS Data. *IGARSS* **2015**, 3365–3368.
168. Saucy, A.; Röösli, M.; Künzli, N.; Tsai, M. Y.; Sieber, C.; Olaniyan, T.; Baatjies, R.; Jeebhay, M.; Davey, M.; Flückiger, B.; Naidoo, R. N.; Dalvie, M. A.; Badpa, M.; de Hoogh, K. Land use regression modelling of outdoor NO2 and PM2.5 concentrations in three low income areas in the western cape province, South Africa. *Int. J. Environ. Res. Public Health* **2018**, *15*, doi:10.3390/ijerph15071452.
169. Lv, Y.; Liu, J.; Yang, T. Nonlinear PLS Integrated with Error-Based LSSVM and Its Application to NO2 Modeling. *Ind. Eng. Chem. Res.* **2012**, *51*, 16092–16100, doi:10.1021/ie3005379.
170. Secretaria del Ambiente de Quito *IAMQ/18*; Quito, 2018;
171. Romero, D. El parque automotor aumenta y complica más la movilidad. *El Comer.* 2017, 1.
172. Todoroski Air Sciences *AIR QUALITY IMPACT ASSESSMENT SANDY POINT QUARRY EPL VARIATION*; Eastwood, 2019;
173. US Department of Commerce, NOAA, E. S. R. L. ESRL Global Monitoring Division - Ozone and Water Vapor Group Available online: https://www.esrl.noaa.gov/gmd/ozwv/surfoz/ (accessed on May 23, 2018).
174. US EPA, O. Ozone Pollution Available online: https://www.epa.gov/ozone-pollution (accessed on May 23, 2018).
175. Monks, P. S.; Archibald, A. T.; Colette, A.; Cooper, O.; Coyle, M.; Derwent, R.; Fowler, D.; Granier, C.; Law, K. S.; Mills, G. E.; Stevenson, D. S.; Tarasova, O.; Thouret, V.; Von Schneidemesser, E.; Sommariva, R.; Wild, O.; Williams, M. L. Tropospheric ozone and its precursors from the urban to the global scale from air quality to short-lived climate forcer. *Atmos. Chem. Phys.* **2015**, *15*, 8889–8973, doi:10.5194/acp-15-8889-2015.
176. WHO (World Health Organization) Health risks of ozone from long-range transboundary air pollution. *J. Chem. Inf. Model.* **2013**, *53*, 1689–1699, doi:10.1017/CBO9781107415324.004.
177. Lee, P.; Saylor, R.; McQueen, J. Air Quality Monitoring and Forecasting. *Atmosphere (Basel).* **2018**, *9*, 89, doi:10.3390/atmos9030089.
178. Chen, L.; Bai, Z.; Kong, S.; Han, B.; You, Y.; Ding, X.; Du, S.; Liu, A. A land use regression for predicting NO2 and PM10 concentrations in different seasons in Tianjin region, China. *J. Environ.*

Sci. **2010**, *22*, 1364–1373, doi:10.1016/S1001-0742(09)60263-1.

179. Zhang, X.; Chu, Y.; Wang, Y.; Zhang, K. Predicting daily PM2.5 concentrations in Texas using high-resolution satellite aerosol optical depth. *Sci. Total Environ.* **2018**, *631–632*, 904–911, doi:10.1016/J.SCITOTENV.2018.02.255.
180. Meng, X.; Chen, L.; Cai, J.; Zou, B.; Wu, C. F.; Fu, Q.; Zhang, Y.; Liu, Y.; Kan, H. A land use regression model for estimating the NO2 concentration in shanghai, China. *Environ. Res.* **2015**, *137*, 308–315, doi:10.1016/j.envres.2015.01.003.
181. Zheng, S.; Zhou, X.; Singh, R.; Wu, Y.; Ye, Y.; Wu, C. The Spatiotemporal Distribution of Air Pollutants and Their Relationship with Land-Use Patterns in Hangzhou City, China. *Atmosphere (Basel).* **2017**, *8*, 110, doi:10.3390/atmos8060110.
182. Braun, D.; Damm, A.; Hein, L.; Petchey, O. L.; Schaepman, M. E. Spatio-temporal trends and trade-offs in ecosystem services: An Earth observation based assessment for Switzerland between 2004 and 2014. *Ecol. Indic.* **2018**, *89*, 828–839, doi:10.1016/J.ECOLIND.2017.10.016.
183. Daac, N. L. P.; Falls, S.; March, S. D. MODIS Land Products Quality Assurance Tutorial : Part- - - 1 How to find , understand , and use the quality assurance information for MODIS land products. **2012**, 1–15.
184. Jia, K.; Liang, S.; Zhang, L.; Wei, X.; Yao, Y.; Xie, X. Forest cover classification using Landsat ETM+ data and time series MODIS NDVI data. *Int. J. Appl. Earth Obs. Geoinf.* **2014**, *33*, 32–38, doi:10.1016/j.jag.2014.04.015.
185. Zhang, J.; Hu, J.; Lian, J.; Fan, Z.; Ouyang, X.; Ye, W. Seeing the forest from drones: Testing the potential of lightweight drones as a tool for long-term forest monitoring. *Biol. Conserv.* **2016**, *198*, 60–69, doi:10.1016/j.biocon.2016.03.027.
186. Mok, K. M.; Yuen, K. V.; Hoi, K. I.; Chao, K. M.; Lopes, D. Predicting ground-level ozone concentrations by adaptive Bayesian model averaging of statistical seasonal models. *Stoch. Environ. Res. Risk Assess.* **2018**, *32*, 1283–1297, doi:10.1007/s00477-017-1473-1.
187. Cazorla, M. Air quality over a populated andean region: Insights from measurements of ozone, NO, and boundary layer depths. *Atmos. Pollut. Res.* **2016**, *7*, 66–74, doi:10.1016/j.apr.2015.07.006.
188. US EPA *Report of the Environment: Ozone Concentrations*; 2014;
189. NASA EOSDIS Remote Sensors Available online: https://earthdata.nasa.gov/user-resources/remote-sensors (accessed on Jul 21, 2018).
190. Sicard, P.; Anav, A.; De Marco, A.; Paoletti, E. Projected global ground-level ozone impacts on vegetation under different emission and climate scenarios. *Atmos. Chem. Phys* **2017**, *17*, 12177–

12196, doi:10.5194/acp-17-12177-2017.

191. USGS *Product guide: Landsat surface reflectance-derived spectral indices*; 2017;

192. Larkin, A.; Geddes, J. A.; Martin, R. V; Xiao, Q.; Liu, Y.; Marshall, J. D.; Brauer, M.; Hystad, P. Global Land Use Regression Model for Nitrogen Dioxide Air Pollution. *Environ. Sci. Technol.* **2017**, *51*, 6957–6964, doi:10.1021/acs.est.7b01148.

193. Gholizadeh, H.; Robeson, S. M. Revisiting empirical ocean-colour algorithms for remote estimation of chlorophyll-a content on a global scale Revisiting empirical ocean-colour algorithms for remote estimation of chlorophyll-a content on a global scale. *Int. J. Remote Sens.* **2016**, *37*, 2682–2705, doi:10.1080/01431161.2016.1183834org/10.1080/01431161.2016.1183834.

194. Ann Becerra, T.; Wilhelm, M.; Olsen, J.; Cockburn, M.; Ritz, B. Ambient Air Pollution and Autism in Los Angeles County, California. *Environ. Health Perspect.* **2013**, doi:10.1289/ehp.1205827.

195. Adam-Poupart, A.; Brand, A.; Fournier, M.; Jerrett, M.; Smargiassi, A. Spatiotemporal Modeling of Ozone Levels in Quebec (Canada): A Comparison of Kriging, Land-Use Regression (LUR), and Combined Bayesian Maximum Entropy–LUR Approaches. *Environ. Health Perspect.* **2014**, doi:10.1289/ehp.1306566.

196. Wolf, K.; Cyrys, J.; Harciníková, T.; Gu, J.; Kusch, T.; Hampel, R.; Schneider, A.; Peters, A. Land use regression modeling of ultrafine particles, ozone, nitrogen oxides and markers of particulate matter pollution in Augsburg, Germany. *Sci. Total Environ.* **2017**, *579*, 1531–1540, doi:10.1016/J.SCITOTENV.2016.11.160.

197. North, G. R. CLIMATE AND CLIMATE CHANGE | Greenhouse Effect. *Encycl. Atmos. Sci.* **2015**, 80–86, doi:10.1016/B978-0-12-382225-3.00470-9.

198. Chi, Y.; Shi, H.; Zheng, W.; Sun, J. Simulating spatial distribution of coastal soil carbon content using a comprehensive land surface factor system based on remote sensing. *Sci. Total Environ.* **2018**, *628–629*, 384–399, doi:10.1016/j.scitotenv.2018.02.052.

199. de Mesnard, L. Pollution models and inverse distance weighting: Some critical remarks. *Comput. Geosci.* **2013**, *52*, 459–469, doi:10.1016/J.CAGEO.2012.11.002.

200. Zheng, C.; Zhao, C.; Li, Y.; Wu, X.; Zhang, K.; Gao, J.; Qiao, Q.; Ren, Y.; Zhang, X.; Chai, F. Spatial and temporal distribution of NO2 and SO2 in Inner Mongolia urban agglomeration obtained from satellite remote sensing and ground observations. *Atmos. Environ.* **2018**, *188*, 50–59, doi:10.1016/J.ATMOSENV.2018.06.029.

201. Ibrahim Sameen, M.; Kubaisy, M. A. Al; Nahhas, F. H.; Ali, A. A.; Othman, N.; Hason, M. M. Sulfur Dioxide (SO2) Monitoring Over Kirkuk City Using Remote Sensing Data. *J. Civ. Environ. Eng.* **2014**, *04*, 1–6, doi:10.4172/2165-784X.1000155.

202. EPA *Inventory of U.S. Greenhouse Gas Emissions and Sinks: 1990-2017*; 2017;

203. Di, Q.; Wang, Y. Y.; Zanobetti, A.; Wang, Y. Y.; Koutrakis, P.; Choirat, C.; Dominici, F.; Schwartz, J. D. Air Pollution and Mortality in the Medicare Population. *N. Engl. J. Med.* **2017**, *376*, 2513–2522, doi:10.1056/NEJMoa1702747.

204. Wulder, M. A.; Loveland, T. R.; Roy, D. P.; Crawford, C. J.; Masek, J. G.; Woodcock, C. E.; Allen, R. G.; Anderson, M. C.; Belward, A. S.; Cohen, W. B.; Dwyer, J.; Erb, A.; Gao, F.; Griffiths, P.; Helder, D.; Hermosilla, T.; Hipple, J. D.; Hostert, P.; Hughes, M. J.; Huntington, J.; Johnson, D. M.; Kennedy, R.; Kilic, A.; Li, Z.; Lymburner, L.; McCorkel, J.; Pahlevan, N.; Scambos, T. A.; Schaaf, C.; Schott, J. R.; Sheng, Y.; Storey, J.; Vermote, E.; Vogelmann, J.; White, J. C.; Wynne, R. H.; Zhu, Z. Current status of Landsat program, science, and applications. *Remote Sens. Environ.* **2019**, *225*, 127–147, doi:10.1016/J.RSE.2019.02.015.

205. Zhai, L.; Sang, H.; Zhang, J.; An, F. Estimating the spatial distribution of $PM_{2.5}$ concentration by integrating geographic data and field measurements. *ISPRS - Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2015**, *XL-7/W4*, 209–213, doi:10.5194/isprsarchives-XL-7-W4-209-2015.

206. Famoso, F.; Wilson, J.; Monforte, P.; Lanzafame, R.; Brusca, S.; Lulla, V. *Measurement and Modeling of Ground-Level Ozone Concentration in Catania, Italy using Biophysical Remote Sensing and GIS*; 2017; Vol. 12;.

207. Anderson, H. R.; Butland, B. K.; van Donkelaar, A.; Brauer, M.; Strachan, D. P.; Clayton, T.; van Dingenen, R.; Amann, M.; Brunekreef, B.; Cohen, A.; Dentener, F.; Lai, C.; Lamsal, L. N.; Martin, R. V.; One, I. P. Satellite-based Estimates of Ambient Air Pollution and Global Variations in Childhood Asthma Prevalence. *Environ. Health Perspect.* **2012**, *120*, 1333–1339, doi:10.1289/ehp.1104724.

208. *Health at a Glance 2017*; Health at a Glance; OECD, 2017; ISBN 9789264280397.

209. Khreis, H.; Nieuwenhuijsen, M.; Khreis, H.; Nieuwenhuijsen, M. J. Traffic-Related Air Pollution and Childhood Asthma: Recent Advances and Remaining Gaps in the Exposure Assessment Methods. *Int. J. Environ. Res. Public Health* **2017**, *14*, 312, doi:10.3390/ijerph14030312.

210. Chang, H.-T.; Wu, C.-D.; Pan, W.-C.; Lung, S.-C. C.; Su, H.-J.; Chang, H.-T.; Wu, C.-D.; Pan, W.-C.; Lung, S.-C. C.; Su, H.-J. Association Between Surrounding Greenness and Schizophrenia: A Taiwanese Cohort Study. *Int. J. Environ. Res. Public Health* **2019**, *16*, 1415, doi:10.3390/ijerph16081415.

211. Sorek-Hamer, M.; Just, A. C.; Kloog, I. Satellite remote sensing in epidemiological studies. *Curr. Opin. Pediatr.* **2016**, *28*, 228–34, doi:10.1097/MOP.0000000000000326.

212. Chan, C.-C.; Chuang, K.-J.; Chen, W.-J.; Chang, W.-T.; Lee, C.-T.; Peng, C.-M. Increasing

cardiopulmonary emergency visits by long-range transported Asian dust storms in Taiwan. *Environ. Res.* **2008**, *106*, 393–400, doi:10.1016/J.ENVRES.2007.09.006.

213. Heo, S.; Bell, M. L. The influence of green space on the short-term effects of particulate matter on hospitalization in the U.S. for 2000–2013. *Environ. Res.* **2019**, *174*, 61–68, doi:10.1016/J.ENVRES.2019.04.019.
214. Jin, J.; Wang, Q.; Jin, J.; Wang, Q. Evaluation of Informative Bands Used in Different PLS Regressions for Estimating Leaf Biochemical Contents from Hyperspectral Reflectance. *Remote Sens.* **2019**, *11*, 197, doi:10.3390/rs11020197.
215. Sobrino, J. A.; Jiménez-Muñoz, J. C.; Paolini, L. Land surface temperature retrieval from LANDSAT TM 5. *Remote Sens. Environ.* **2004**, *90*, 434–440, doi:10.1016/j.rse.2004.02.003.
216. Chen, S.; Goo, Y.-J. J.; Shen, Z.-D. A hybrid approach of stepwise regression, logistic regression, support vector machine, and decision tree for forecasting fraudulent financial statements. *ScientificWorldJournal.* **2014**, *2014*, 968712, doi:10.1155/2014/968712.
217. Zhang, Z. Variable selection with stepwise and best subset approaches. *Ann. Transl. Med.* **2016**, *4*, 136, doi:10.21037/atm.2016.03.35.
218. Olive, D. J. *Linear Regression*; Springer International Publishing, 2017; ISBN 9783319552521.
219. Habermann, M.; Billger, M.; Haeger-Eugensson, M. Land use Regression as Method to Model Air Pollution. Previous Results for Gothenburg/Sweden. *Procedia Eng.* **2015**, *115*, 21–28, doi:10.1016/J.PROENG.2015.07.350.
220. Fan, J.; Li, S.; Fan, C.; Bai, Z.; Yang, K. The impact of PM2.5 on asthma emergency department visits: a systematic review and meta-analysis. *Environ. Sci. Pollut. Res.* **2016**, *23*, 843–850, doi:10.1007/s11356-015-5321-x.
221. Humans, I. W. G. on the E. of C. R. to; Cancer, I. A. for R. on *Outdoor Air Pollution*; IARC Monographs on the Evaluat; International Agency for Research on Cancer, World Health Organization, 2016; ISBN 9789283201472.
222. D'Amato, G.; Cecchi, L.; D'Amato, M.; Annesi-Maesano, I. Climate change and respiratory diseases. *Eur. Respir. Rev.* **2014**, *23*, 161–9, doi:10.1183/09059180.00001714.
223. Arundel, A. V; Sterling, E. M.; Biggin, J. H.; Sterling, T. D. Indirect health effects of relative humidity in indoor environments. *Environ. Health Perspect.* **1986**, *65*, 351, doi:10.1289/EHP.8665351.
224. Hadjimitsis, D. G. Aerosol optical thickness (AOT) retrieval over land using satellite image-based algorithm. *Air Qual. Atmos. Heal.* **2009**, *2*, 89–97, doi:10.1007/s11869-009-0036-0.
225. Ayehu, G.; Tadesse, T.; Gessesse, B.; Yigrem, Y.; Ayehu, G.; Tadesse, T.; Gessesse, B.; Yigrem,

Y. Soil Moisture Monitoring Using Remote Sensing Data and a Stepwise-Cluster Prediction Model: The Case of Upper Blue Nile Basin, Ethiopia. *Remote Sens.* **2019**, *11*, 125, doi:10.3390/rs11020125.

226. Hastie, T.; Tibshirani, R.; Friedman, J.; Franklin, J. The elements of statistical learning: data mining, inference and prediction. *Math. Intell.* **2005**, *27*, 83–85.

Annex I

Remote Sensing Applications: Society and Environment 13 (2019) 257–274

Contents lists available at ScienceDirect

Remote Sensing Applications: Society and Environment

journal homepage: www.elsevier.com/locate/rsase

Improving NDVI by removing cirrus clouds with optical remote sensing data from Landsat-8 – A case study in Quito, Ecuador

Cesar I. Alvarez-Mendoza[a,b,*], Ana Teodoro[a,c], Lenin Ramirez-Cando[b]

[a] University of Porto, Department of Geosciencies, Environment and Land Planning, Faculty of Sciences, Rua Campo Alegre 687, Porto 4169-007, Portugal
[b] Universidad Politécnica Salesiana, Grupo de Investigación Ambiental en el Desarrollo Sustentable, Carrera de Ingeniería Ambiental, Quito, Ecuador
[c] Earth Sciences Institute (ICT), Pole of the FCUP, University of Porto, Porto, Portugal

ARTICLE INFO

Keywords:
Cloud removal
Optical remote sensing
Landsat-8 OLI
Quito
NDVI

ABSTRACT

The Andean region has a high cloud density throughout the year. The use of optical remote sensing data in the computation of environmental indices of this region has been hampered by the presence of clouds. To maximize accuracy in the computation of several environmental indices including the normalized difference vegetation index (NDVI), we compared the performance of two algorithms in removing clouds in Landsat-8 Operational Land Imager (OLI) data of a high-elevation area. The study area was Quito, Ecuador, which is a city located close to the equator and in a high-elevation area crossed by the Andes Mountains. The first algorithm was the automatic cloud removal method (ACRM), which employs a linear regression between the different spectral bands and the cirrus band. The second algorithm was independent component analysis (ICA), which considers the noise (clouds) as part of independent components applied over the study area. These methods were evaluated based on several images from different years with different cloud conditions. The results indicate that neither algorithm is effective over this region for the removal of clouds or for NDVI computation. However, after improving ACRM, the NDVI computed using ACRM showed a better correlation than ICA with the MODIS NDVI product.

1. Introduction

Optical remote sensing (ORS) data have the major advantage of providing synoptic and frequent overviews of the Earth's surface, but the distribution of ground-based measurements is scarce in some parts of the world. ORS data include visible (VIS), short-infrared (SWIR), and thermal infrared (TIR) regions of the electromagnetic spectrum (Lillesand et al., 2015).

Regions with a high cloud density during most of the year, such as the Brazilian Amazon (Rees, 2012; Ju and Roy, 2008; Asner, 2001) and the Andean region (Fernández et al., 2015), are particularly challenging for ORS, especially in terms of the computation of the environmental indices, such as normalized difference vegetation index (NDVI) (Weier and Herring, 2000; Rajitha et al., 2015). Several studies on cloud density have been conducted based on Landsat data (Rees, 2012; Asner, 2001; Ju and Roy, 2008). Richter et al. (2011) takes the spectral/spatial characteristics of Sentinel-2 as a template for instruments with similar properties as Sentinel-2 to investigate the relevant cirrus effects. Shen et al. (2014) proposed a method based on the classic homomorphic filter executed in the frequency domain to thin cloud removal for visible remote sensing images. Gao and Li (2017) propose an empirical technique for the removal of thin cirrus scattering effects in OLI visible near infrared and shortwave IR spectral regions. In the work of Lv et al. (2018), the top-of-atmosphere reflectance of thin clouds is modeled using the empirical relationships of the deep blue and blue bands of Landsat-8 OLI.

The Landsat program has provided calibrated and high-resolution spatial data of the Earth's surface for more than 45 years. Landsat-8, launched in February 2013, is the latest satellite in a continuous series of land remote sensing satellites that began in 1972. Landsat-8 has provided data to support several fields and research topics, such as agriculture, forestry, geology, land use, air contamination (USGS, 2013), and the removal of clouds in remote sensing images (Hashim et al., 2014; Pour and Hashim, 2017; Lv et al., 2016; Cheng et al., 2014;

* Correspondence author at: University of Porto, Department of Geosciencies, Environment and Land Planning, Faculty of Sciences, Rua Campo Alegre 687, Porto 4169-007, Portugal.
E-mail addresses: calvarezm@ups.edu.ec, calvarezm@ups.edu.ec (C.I. Alvarez-Mendoza), amteodor@fc.up.pt (A. Teodoro), lramirez@ups.edu.ec (L. Ramirez-Cando).

https://doi.org/10.1016/j.rsase.2018.11.008
Received 7 June 2018; Received in revised form 24 October 2018; Accepted 13 November 2018
Available online 16 November 2018

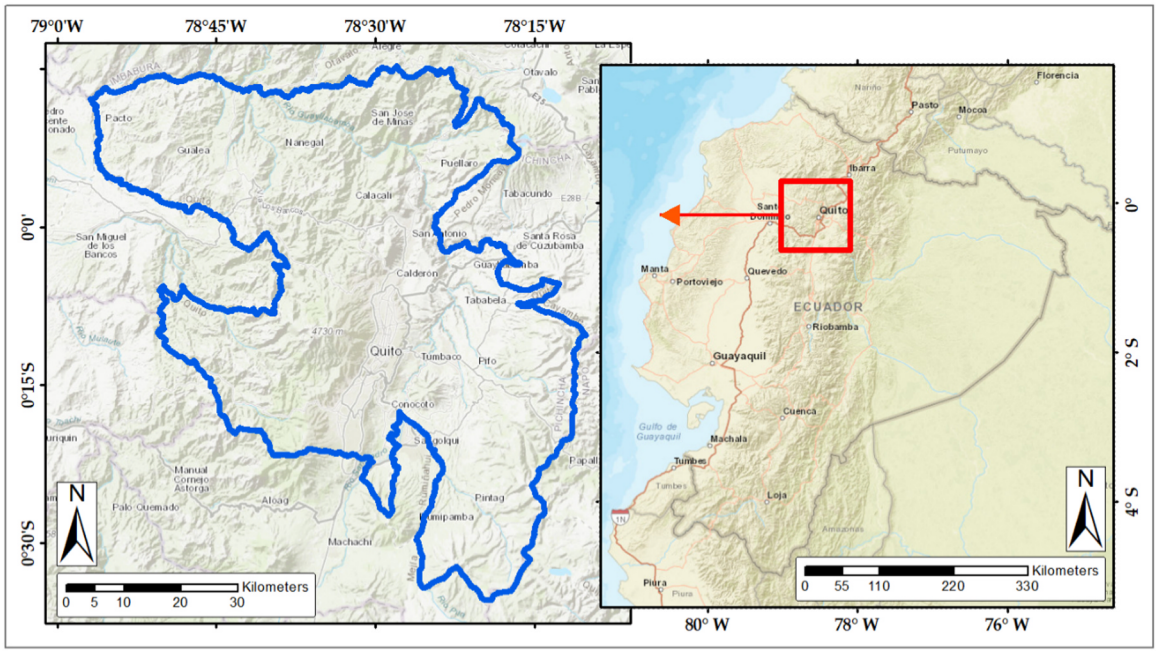


Fig. 1. Quito metropolitan area.

Gao and Li, 2012; Shen et al., 2015a, 2015b; Huadong et al., 2009; Zhu and Woodcock, 2012; Xu et al., 2014). Landsat-8 includes two sensors: the Operational Land Imager (OLI), which is divided into nine bands with a spatial resolution of 30 m, and the Thermal Infrared Sensor (TIRS) instrument, which is divided into two bands with a native spatial resolution of 100 m. The OLI bands include a cirrus band (B9). Cirrus clouds are high-altitude clouds in the atmosphere and are mainly composed of miniscule ice crystals (Stephens, 2005). They are strong reflectors of radiation at a wavelength of 1.38 μm (Department of the Interior U.S. Geological Survey, 2016). Cirrus clouds have a significant number of thin, non-spherical ice crystals that can absorb sunlight and attenuate the pixel values of surface reflectance in remote sensing (Gao et al., 1998). Additionally, cirrus clouds limit the accuracy in the computation of environmental indices. Thus, it is crucial to remove them (Rajitha et al., 2015).

The purpose of this work is to develop an approach to remove clouds and noise in optical remote sensing data without losing surface pixel accuracy in order to compute environmental indices, such as NDVI. Several methods have been tested to remove clouds considering Landsat-8 data in different places around the world with satisfactory results. Some of these methods used a reference Landsat-8 image to patch the cloudy area (Cheng et al., 2014; Lin et al., 2014; Lv et al., 2016), or combine Landsat-8 with other sensors (Wu et al., 2016), or work with the Landsat-8 cirrus band (B9) (Shen et al., 2015a, 2015b; Xu et al., 2014). All these studies were conducted in low elevation regions and in no tropical areas. Both parameters can have an effect over cirrus clouds (He et al., 2013), considering that these clouds can form at any altitude between 5.0 km and 14 km above sea level. In the tropical regions, cirrus clouds cover around 70% of the region's surface area.

In this work, to remove cirrus clouds over an area in the Andean region (Quito, Ecuador) considering the Landsat-8 cirrus band (B9), two methods were evaluated: the automatic cloud removal method (ACRM) and independent component analysis (ICA). ACRM was first tested on images of Sydney, Australia (Xu et al., 2014). The algorithm applies a linear regression between each multispectral band and the cirrus band (B9), evaluates the coefficient of determination (R^2) and slope in some areas, and generalizes them for the entire image (Xu et al., 2014). In order to remove clouds, the algorithm uses the area with the highest R^2 to extrapolate values for the entire image. In ICA, independent components (ICs) are separated, and one of them is the component that storing the thin clouds (Hyvärinen and Oja, 2000). This algorithm was tested on Landsat-8 images of a low elevation region (North Carolina, USA), and the results were satisfactory (Shen et al., 2015a, 2015b). The performance of the two methods in removing clouds and their efficiency in future computation of environmental indices such as NDVI are evaluated based on the same image.

2. Materials and methods

2.1. Study area and dataset

2.1.1. Study Area

The study area is Quito, the capital of Ecuador (Fig. 1). The equator line crosses the city in the north part. The Quito latitude ranges between 0°30′S to 0°10′N and its longitude ranges between 78°10′W to 78°40′W. Quito has a high elevation of approximately 2800 m. The cloud density over the city is considerable, all over the year. Quito has only one dry season and one wet season, considering that it is a tropical zone and is influenced by the Andes Mountains. In 2015, the mean minimum and maximum temperatures were approximately 9.0 °C and 25.4 °C, respectively, with a high precipitation of approximately 1126 mm (Instituto Nacional de Meteorología e Hidrología, 2016). The geology of northeastern Ecuador and present-day physical processes related to geology are greatly influenced by the tectonic mechanisms responsible for the development of the Andes Mountains. Both geology and active physical processes (landsliding, volcanism, erosion, weathering) are complex and varied (Baldock, 1982).

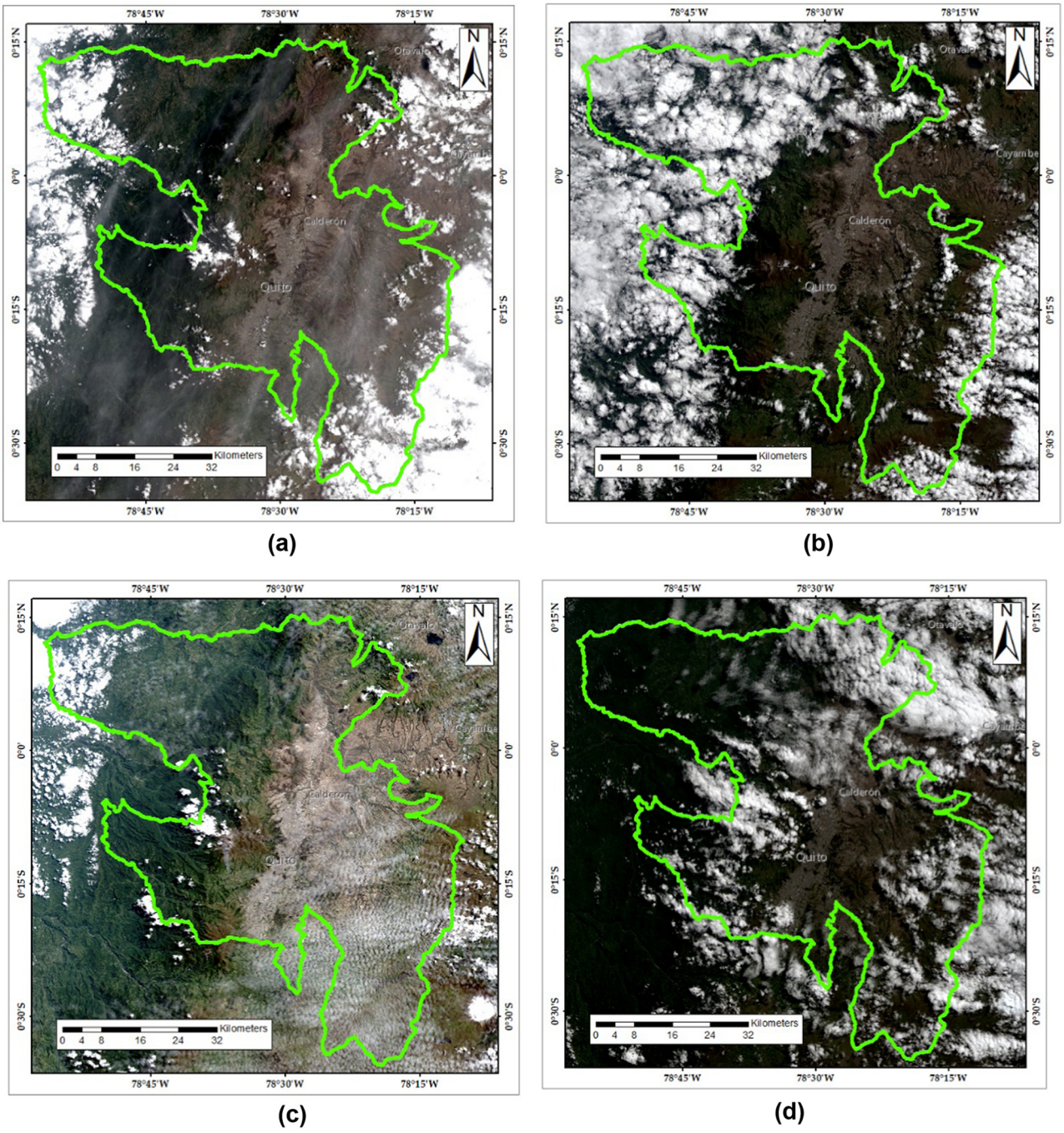


Fig. 2. Landsat-8 Images from Quito Metropolitan Area (Path: 10; Row: 60): (**a**) Image from 2013/10/11; (**b**) Image from 2013/07/07; (**c**) Image from 2014/07/26; (**d**) Image from 2015/07/13; (**e**) Image from 2015/08/30; (**f**) Image from 2016/02/06; (**g**) Image from 2016/10/19; (**h**) Image from 2013/06/21 (Reference image to ICA evaluation).

2.1.2. Dataset

In this study, ten Landsat-8 L1T images were processed to evaluate and improve the two methods to remove clouds. Seven images of Quito, Ecuador (Path 10; Row 60) from different years (Fig. 2); one image of Pedernales, Ecuador (Path 11; Row 60), which is a coastal region with characteristics similar to those of Sydney; and the image of Sydney, Australia (Path 89; Row 83) used in (Xu et al., 2014) were considered.

Images at the L1T processing level were considered because they take advantage of geometric and radiometric corrections (Department of the Interior U.S. Geological Survey, 2016). Moreover, the MODIS

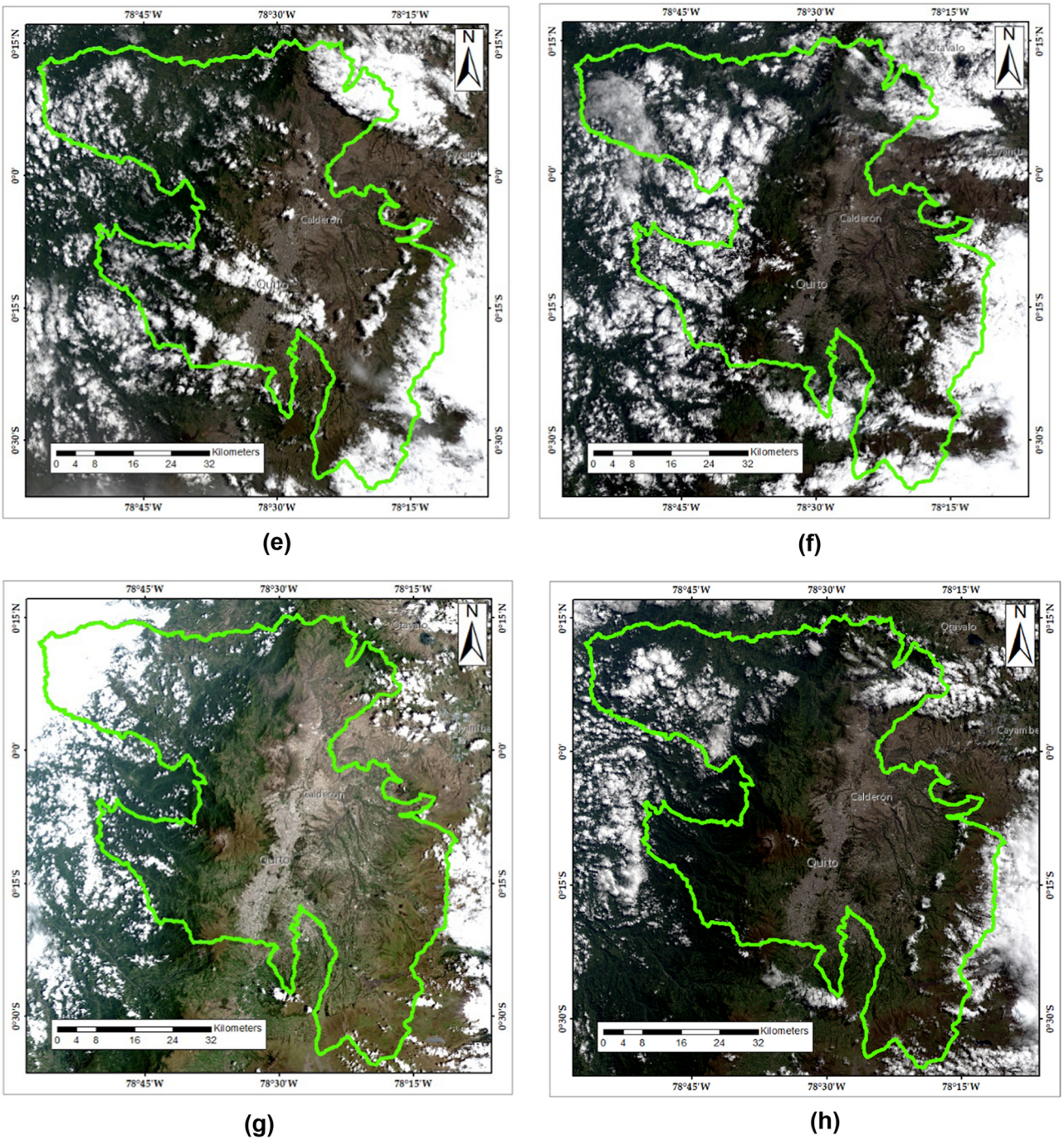


Fig. 2. (*continued*)

Table 1
Characteristics of datasets used in this study.

Sensor	Product	Spatial Resolution	Temporal resolution	Bands/Products
Landsat-8	L1T	30 m	16 days	Coastal aerosol, blue, green, red, near infrared, SWIR 1 and SWIR 2, Cirrus, Thermal Infrared 1, Thermal Infrared 2
MODIS	MOD13Q1	250 m	16 days	NDVI/EVI Values

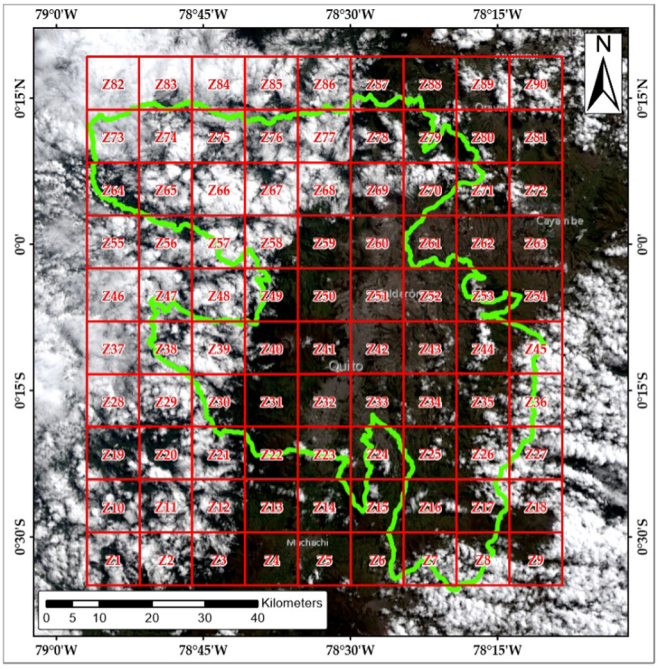


Fig. 3. Input regions considered to test the ACRM algorithm.

MOD13Q1 product (tiles H10V08 and H10V09) for the study area was also used in order to compare the results obtained in the computation of NDVI (further details in Section 3.4) (Table 1).

2.2. Methodology

Two methods to remove clouds, ACRM and ICA, were evaluated in this work for Landsat-8 images and the corresponding cirrus band (B9). Most of the processing steps were implemented in R programming language (R Core Team, 2016) and its associated packages: raster version 2.5–8 (Hijmans, 2016), rgdal version 1.1 (Bivand et al., 2016), and gdalutilities version 2.0.1.7 (Greenberg and Mattiuzzi, 2015). Furthermore, ENVI® and ERDAS® software were used to perform some image processing tasks.

2.2.1. Automatic Cloud Removal Method (ACRM)

ACRM attempts to obtain clean pixel data from each digital number *DN* recorded at each OLI multispectral band $i = 1, 2, 3, 4, 5, 6, 7$. *DN* contains clean pixel data plus contaminated data at the location (u, v). Contaminated data are affected by clouds (Xu et al., 2014). The model can be expressed as follows:

$$DN(u, v) = x_i^f(u, v) + x_i^c(u, v), \quad i = 1, 2, 3, 4, 5, 6, 7, \tag{1}$$

where $x_i^f(u, v)$ is the clean cloud-free pixel from each of bands 1–7 and $x_i^c(u, v)$ is the cirrus cloud pixel from each of bands 1–7 obtained with band 9. Eq. (1) results from the strong linear relationship between the bands found in (Ji, 2008), where $x_i^c(u, v)$ is linearly related to the

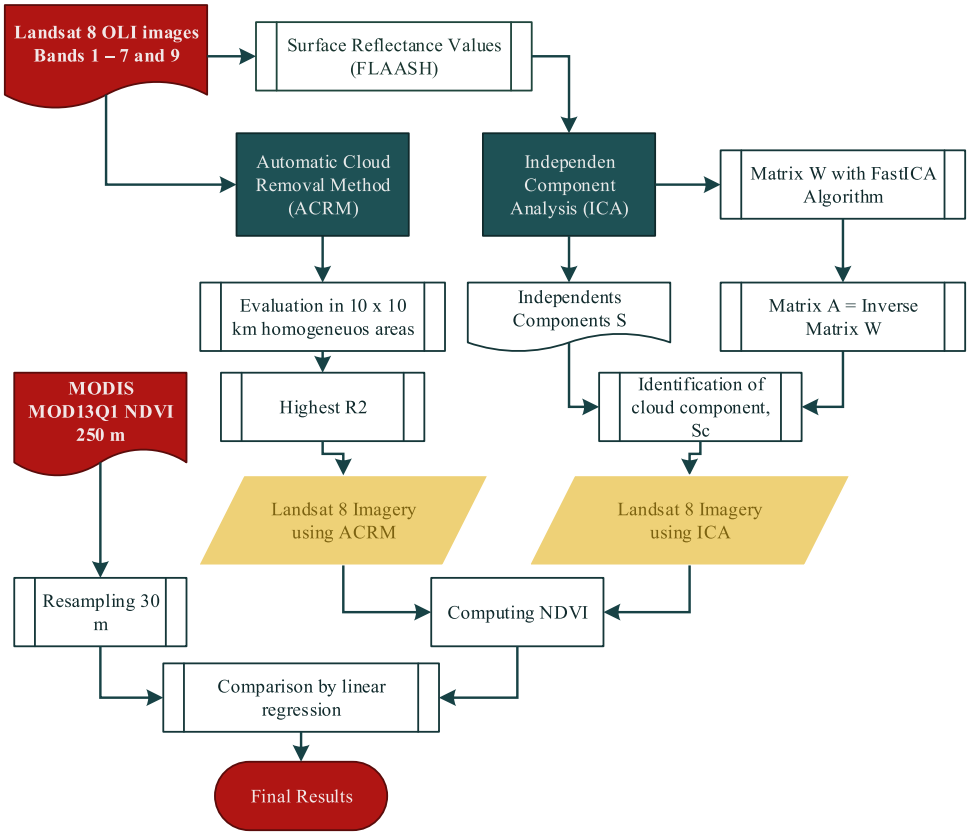


Fig. 4. Flowchart of the methodology adopted to perform a comparison between ACRM and ICA algorithms.

Table 2
Linear regression results between bands 1–7 and 9 in the Quito study area for different dates.

Band	R^2	Slope (α)	R^2	Slope (α)	R^2	Slope (α)	R^2	Slope (α)
	Quito (11/10/2013)		Quito (07/26/2014)		Quito (07/13/2015)		Quito (02/06/2016)	
B2	0.96	0.05	0.93	0.02	0.95	0.03	0.95	0.03
B3	0.96	0.05	0.93	0.02	0.95	0.03	0.95	0.03
B4	0.96	0.05	0.93	0.02	0.95	0.02	0.95	0.02
B5	0.88	0.02	0.85	0.01	0.91	0.02	0.85	0.01
B6	0.85	0.02	0.89	0.17	0.88	0.02	0.89	0.03
B7	0.86	0.02	0.88	0.02	0.87	0.02	0.88	0.02
	Quito (07/07/2013)		Quito (08/30/2015)		Quito (10/19/2016)		Quito (21/06/2013)	
B2	0.96	0.05	0.93	0.02	0.97	0.03	0.95	0.03
B3	0.96	0.06	0.93	0.02	0.97	0.03	0.95	0.03
B4	0.95	0.05	0.93	0.02	0.97	0.02	0.95	0.02
B5	0.85	0.03	0.85	0.01	0.95	0.02	0.85	0.01
B6	0.90	0.06	0.89	0.17	0.92	0.02	0.89	0.03
B7	0.89	0.06	0.88	0.02	0.89	0.03	0.88	0.02

DN recorded in the cirrus band $c(u, v)$ as follows:

$$x_i^c(u, v) = \propto_i [c(u, v) - min\{c(u, v)\}]. \quad (2)$$

The aim is to obtain the slope $\propto_i$ for each band, considering a linear relationship between each multispectral band and band 9 in a homogenous area. Two approaches can be considered to determine this homogenous area. The first approach is a photo-interpretation to find this area by taking, for example, water bodies that have a near-zero pixel value over the near-infrared (NIR) band. However, this approach cannot be used for images that do not contain water bodies. The second approach is to use random areas of a constant size covering the entire region or zones with a specific land use. In this study, we considered the second approach of finding random areas with a size of 10 × 10 km^2, covering the entire study area (Fig. 3). Smaller regions (250 m * 250 m) were also tested, but the results were identical.

By combining Eq. (1) with Eq. (2), $x_i^f(u, v)$ can be estimated as follows:

$$x_i^f(u, v) = DN(u, v) - \propto_i [c(u, v) - min\{c(u, v)\}] \quad (3)$$

2.2.2. Independent Component Analysis (ICA)

ICA is a method for finding underlying factors or components from multivariate (multidimensional) statistical data (Hyvärinen et al., 2001). The relationship is represented as follows:

$$\mathbf{X} = \mathbf{AS} \quad (4)$$

where **S** is a random vector containing the independent source signal or independent components (IC) with elements $s_1, s_2, \ldots,$ and s_n. A is the "mixing" square matrix having elements a_{ij}. **X** is the observed signal (mixed) having elements $x_1, x_2, \ldots,$ and x_n.

In Eq. (4), **X** represents surface reflectance data from each of bands 1–7 and pixel cirrus data from band 9. The surface reflectance data were obtained by applying atmospheric correction with the fast line-of-sight atmospheric analysis of hypercubes (FLAASH) algorithm (ENVI, 2009; Allred et al., 1994). FLAASH works as a physical method to obtain surface reflectance, and it allows us to describe the shape of the signatures (Mandanici et al., 2015) in ENVI software. The column vector ***s*** represents ICs and matrix **A** represents the linear transformation. Both ***s*** and **A** are unknown.

In some studies, ICA is used to separate some parts of satellite images by considering their bands as ICs. The algorithm achieves cloud removal by considering that each IC is a linear mixture of bands 1–7 and 9. Band 9 is used to delineate the cloud component in the IC (Huadong et al., 2009; Shen et al., 2015a).

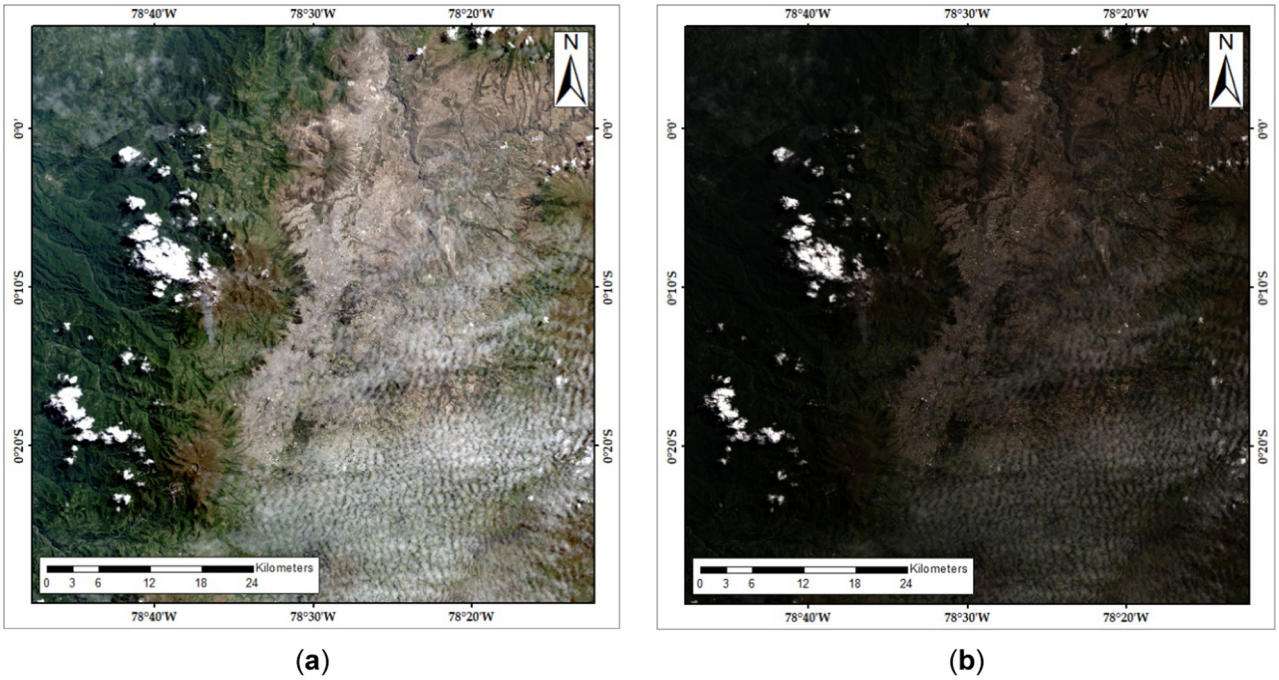


Fig. 5. Landsat-8 Images from Quito Metropolitan Area (Path: 10 Row: 60): Image from 2014/07/26 (**a**) Original Image; (**b**) Image applied ACRM.

Table 3
Linear regression results between bands 1–7 and 9 in the other evaluated zones.

	Sydney (2013/10/04)		Pedernales (2016/05/13)	
Band	**R^2**	**Slope (α)**	**R^2**	**Slope (α)**
B2	0.97	1.70	0.67	0.69
B3	0.99	1.63	0.68	0.68
B4	0.98	1.68	0.67	0.62
B5	0.98	1.74	0.67	0.52
B6	0.99	1.11	0.63	0.44
B7	0.98	1.02	0.53	0.58

ICA works with a non-Gaussian distribution, where ICs (surface reflectance and pixel cloud data) are not normally distributed, because various surface types and cloud types produce different reflectance values. The robust FastICA algorithm can be applied to estimate an unmixing matrix **W**, which is the inverse of mixing matrix **A** (Hyvärinen and Oja, 2000). The source vector ***s*** can be obtained by inverting Eq. (4) as follows:

$$\boldsymbol{s} = \mathbf{A}^{-1}\mathbf{X}. \tag{5}$$

Band 9 (cirrus band, which is a part of **X**) is considered the sum of eight products (bands 1–7 and 9) for each IC: the product of each source vector with its coefficients in **A**. Eq. (6), derived from Eq. (4), allows us to obtain the cloud pixel value $\mathbf{x_{17}}$ as follows:

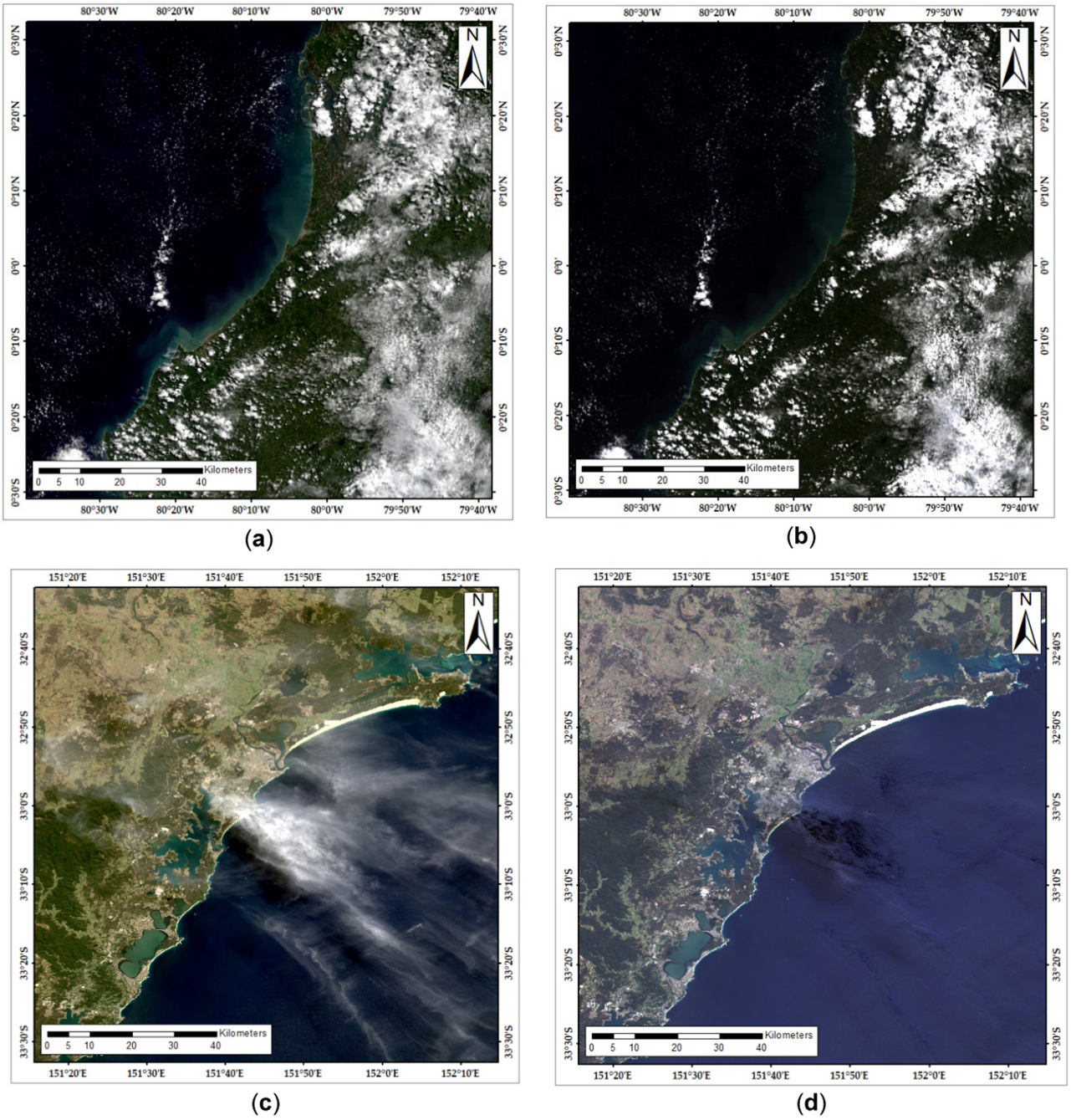


Fig. 6. Landsat 8 OLI images (**a**) Original image from Pedernales; (**b**) Image after applied ACRM in Pedernales; (**c**) Original image from Sidney; (**d**) Image after applied ACRM in Sidney.

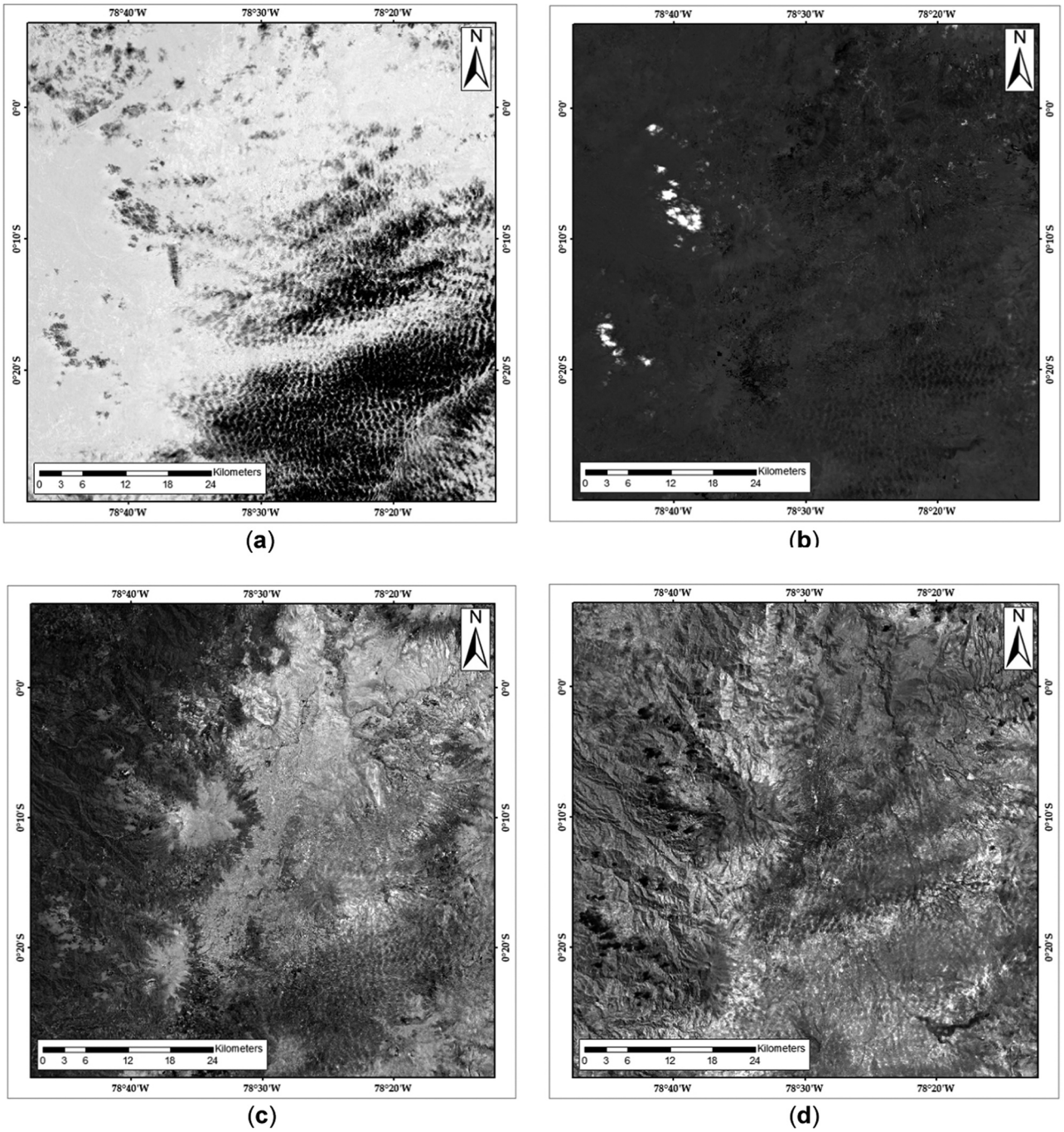


Fig. 7. (a–h) are first, second, …, and eighth independent components, respectively.,.

$$\mathbf{x_{1-7}} = a_{1-7}\boldsymbol{s_c}, \tag{6}$$

Where a_{1-7} is the coefficient of $\boldsymbol{s_c}$ in matrix **A** corresponding to the reflectance data of bands 1–7. The largest factor in the row corresponding to band 9 of **A** determines the $\boldsymbol{s_c}$ to be used to obtain the cloud reflectance data $\boldsymbol{x_c}$. The final reflectance-free data x_f is obtained by subtracting the original reflectance data from each band $\boldsymbol{x_o}$ by the cloud reflectance data from each band $\boldsymbol{x_c}$ (Eq. (7)).

$$x_f = x_o - x_c. \tag{7}$$

2.2.3. Normalized Difference Vegetation Index (NDVI)

NDVI is an index that allows to obtain information about the greenest vegetation considering red and NIR bands of a sensor (Tucker,

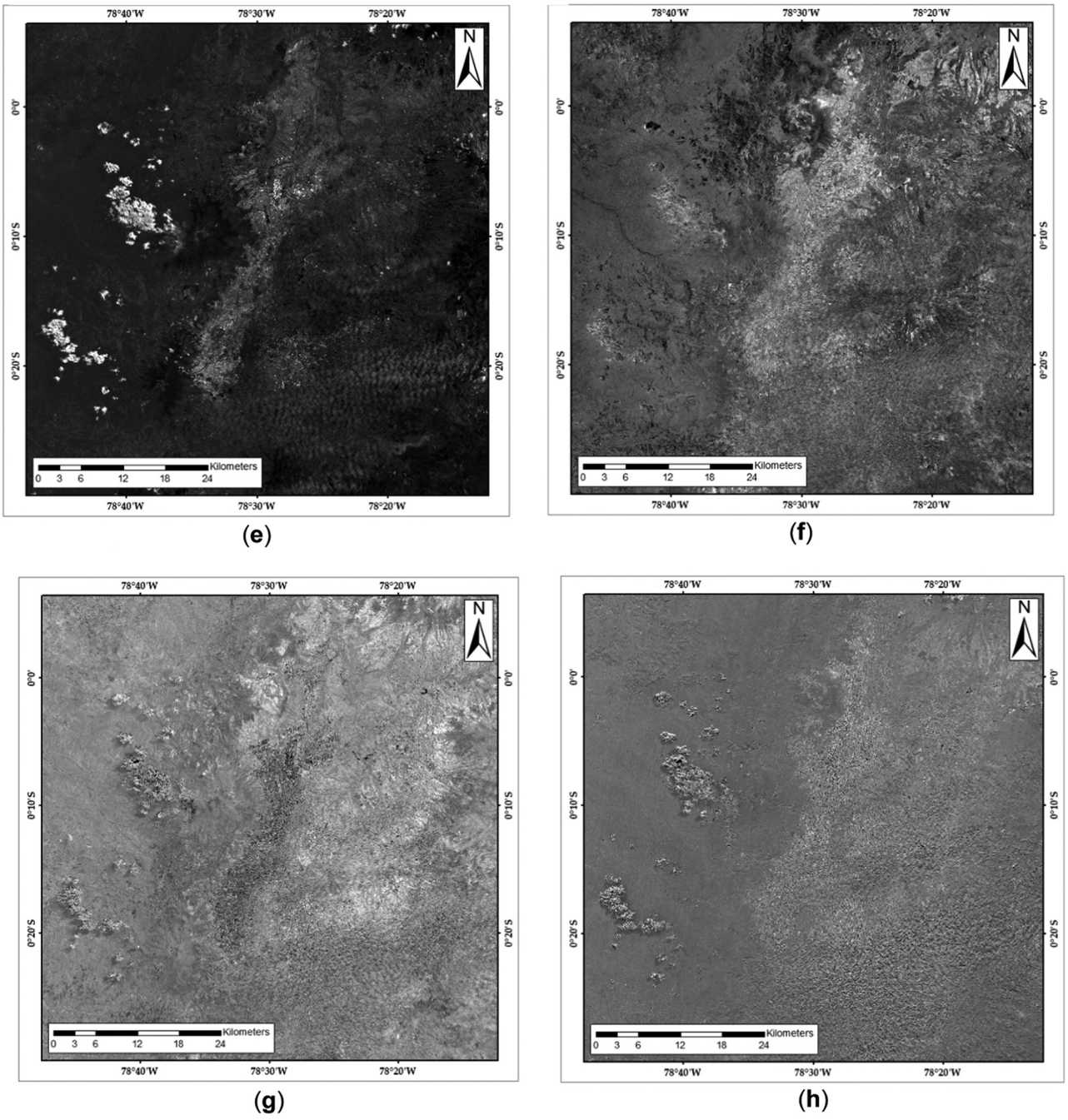


Fig. 7. (*continued*)

Table 4
Coefficients (×10–2) of A.

Band	S1	S2	S3	S4	S5	S6	S7	S8
B1	4.719	0.678	0.653	9.672	1.731	1.818	1.308	0.207
B2	4.613	0.939	0.494	9.192	1.628	1.661	1.722	0.372
B3	4.537	0.802	1.153	8.826	1.645	1.644	2.201	1.149
B4	4.487	0.696	0.851	8.954	1.493	1.692	3.413	1.006
B5	2.824	0.475	0.524	6.962	1.743	1.148	−1.815	7.568
B6	0.236	0.764	1.266	7.093	1.508	1.632	3.497	3.671
B7	0.256	0.901	1.214	6.417	−0.022	1.794	3.746	1.656
B9	−0.021	−0.023	0.018	−0.152	0.984	4.011	0.617	0.108

1979). In the case of Landsat-8 OLI, NDVI is calculated using bands 4 (red band) and 5 (NIR band). The NDVI in a Landsat-8 OLI image is computed as follows (Eq. (8)):

$$NDVI = (B5 - B4)/(B5 + B4) \quad (8)$$

NDVI is one of the most commonly used remote sensing vegetation indices (Roy et al., 2016; Mishra and Mainali, 2017), and it is considered an environmental index owing to its strong relationship with the land surface (e.g., surface temperature, vegetation cover, land use) and meteorological data (e.g., temperature, humidity) (Kuenzer et al., 2015). Moreover, NDVI is used to validate and compare results between

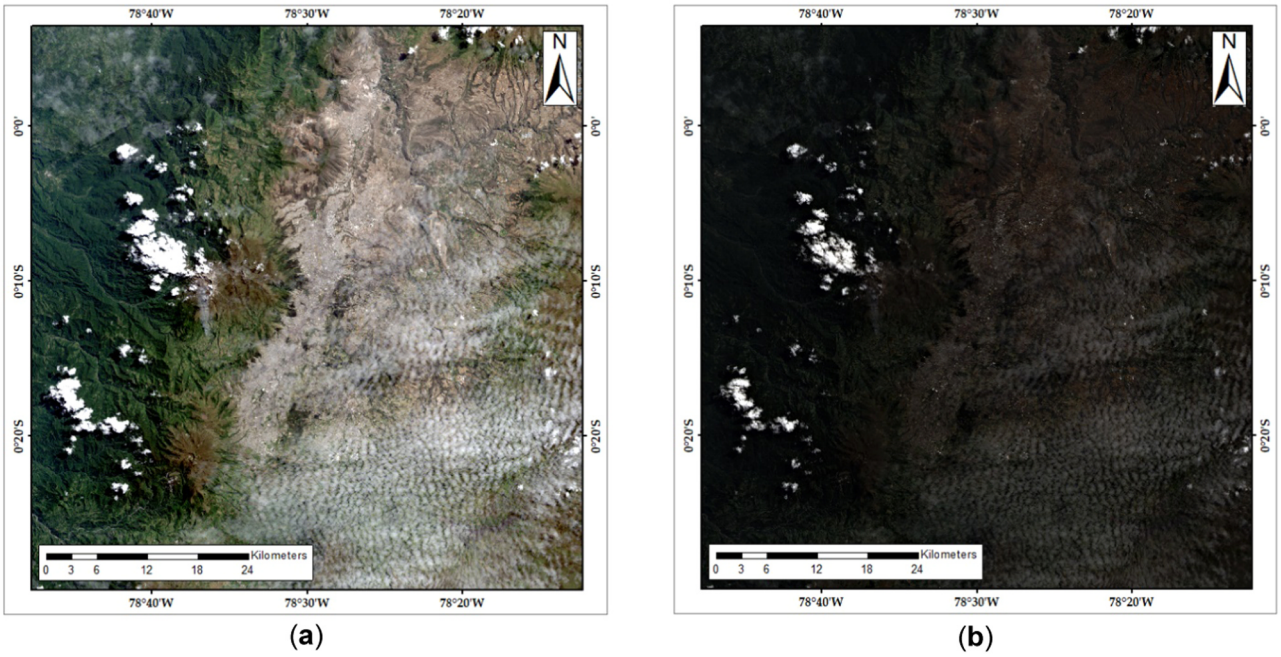


Fig. 8. Landsat-8 Images from Quito Metropolitan Area (Path: 10 Row: 60): Image from 2014/07/26 (**a**) Original Image; (**b**) Image after applied ICA.

sensors by considering future environmental applications (Zambrano et al., 2016).

2.2.4. Evaluation and Validation

In order to validate the efficiency of ACRM and ICA cloud removal methods in the computation of environmental indices, the NDVI was computed in the original Landsat-8 images after applying both algorithms. Then, the images were compared with a MODIS NDVI product resampled to a spatial resolution of 30 m, assuming a similar period of Landsat-8 data used. A MOD13Q1 product (NDVI 16-Day L3 Global 250 m version 6) was used as reference data, considering that MODIS is a ready-to-use product (Huete et al., 2002; Solano et al., 2010) and is evaluated in vegetation phenology. The validation was tested in a small area where cirrus clouds are present, which allowed us to evaluate the performance of the algorithms to remove clouds and to estimate environmental indices. The methodology adopted in this work is presented in the flowchart shown in Fig. 4.

3. Results

3.1. Cloud removal using ACRM

The ACRM algorithm was applied to ten images considered in this study. The code was programmed in R Studio with the raster package. The main objective was to obtain the best correlation (R^2) between bands 1–7 and band 9 in selected areas of the images with cirrus clouds.

The first step was to choose the zones to evaluate the algorithm in a geographic information system (GIS) covering the entire study area in Quito. These areas, called zones (Z), are 10 km × 10 km regular grids covering the study area (Fig. 3). Subsequently, the algorithm was applied, and the best-fit regions with the best R^2 coefficients between each multispectral band (1 − 7) and band 9 (Table 2) were evaluated.

Table 1 lists the highest R^2 coefficients obtained in the application of the algorithm, considering only values higher than 0.85. Slope values are lower than 0.18. These results are shown in Fig. 5 (see Section 4.3).

ACRM was also tested considering an image from Pedernales and an image from Sydney (Table 3). In Pedernales, the R^2 coefficients had values lower than 0.68. Better results were obtained over Sydney with higher R^2 coefficients (higher than 0.97). To corroborate the results of R^2 coefficients (Fig. 6), we confirmed that the image of Pedernales is practically unchanged by the algorithm, while the algorithm removes all the clouds in the image of Sydney.

3.2. Cloud removal considering ICA

The ICA algorithm was applied only to the Quito image from 26/07/2014, which shows clouds over the study area. Different software were used (R Studio, ENVI, ERDAS) to obtain the different parameters showed in the Eq. (4). The principal inputs to the algorithm were the surface reflectance data of multispectral bands (calculated with FLAASH correction from ENVI) and the DN from band 9. Furthermore, the IC for the selected image was obtained in ENVI software with the FastICA algorithm (Hyvärinen and Oja, 2000) (Fig. 7). The matrix **A** from Eq. (6) was obtained using the ICA algorithm in ERDAS software (Table 4), and s_c was selected as s_6, which had the high absolute value of 4.011×10^{-2} in the row of band 9. Then, to obtain the input data for Eq. (7), the product of the coefficient in the column for each band at s_6 with each IC was used. The results are shown in Fig. 8. Again, as in ACRM, the result was not satisfactory in comparison with the original image (see Section 4.3).

Moreover, to corroborate that the application of the ICA algorithm does not provide satisfactory results for Quito, some scatterplots were computed with respect to a cloud-free reference image (Fig. 9). The scatterplots show a linear correlation between the reference image

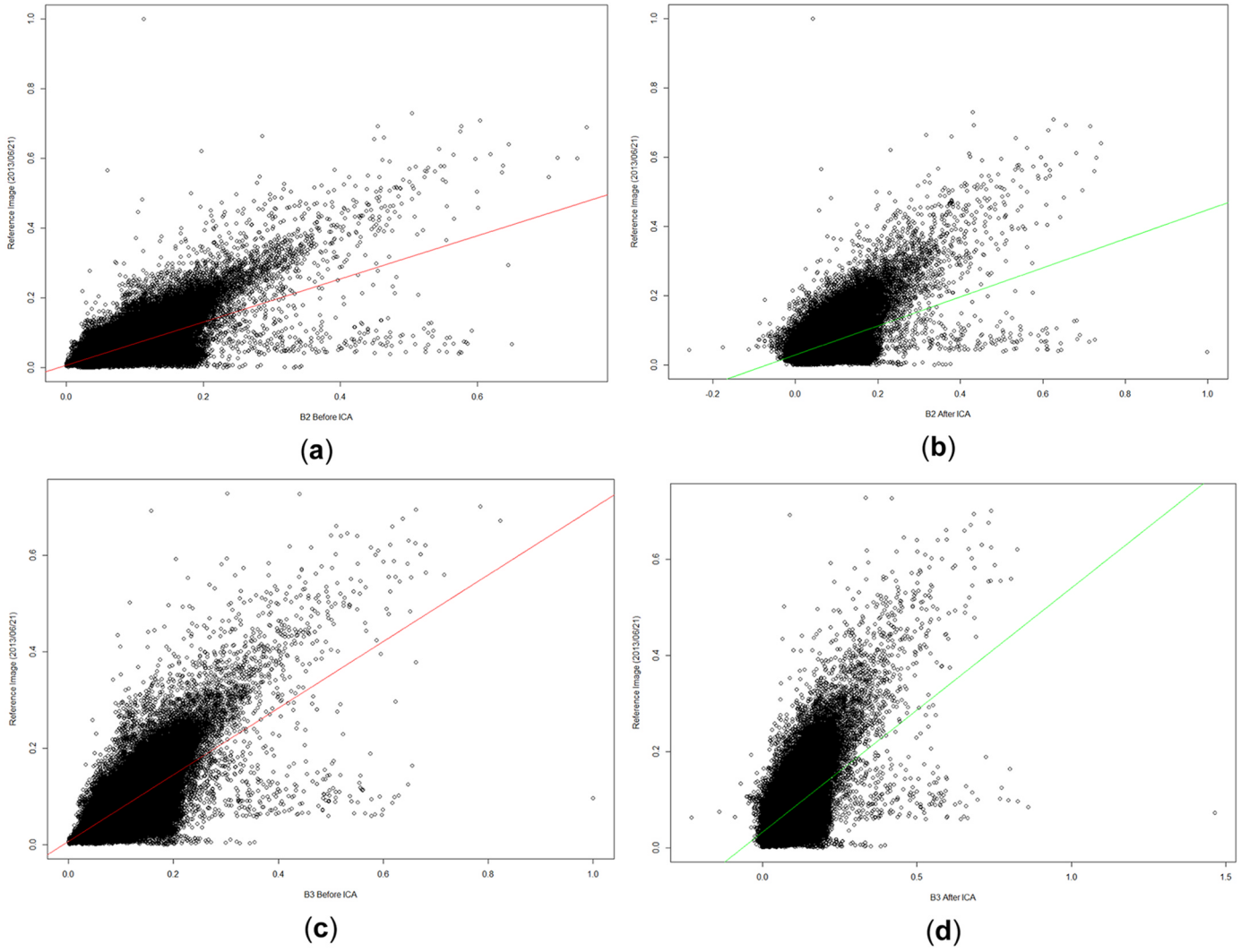


Fig. 9. Scatterplots of bands 2–5. **(a, c, e, g)** Left an image before ICA algorithm implementation vs. reference image. **(b, d, f, h)** Right image considers ICA algorithm implementation vs. Reference image. Reference image is from June 21, 2013 to evaluate ICA (Fig. 2h).

(Fig. 2h) and the images with and without ICA correction (Table 5), which indicates that the ICA algorithm does not work properly for Quito.

As indicated in Table 4, if ICA is applied, the algorithm changes the surface reflectance values; in comparison with a cloud-free image, the correlation decreases.

3.3. Validation – NDVI computation

As mentioned previously, one of the main objectives of the cloud removal in high-altitude areas is to obtain a better accuracy in the computation of environmental indices, such as NDVI. Therefore, in the process of validation of the proposed algorithms, the NDVI values for a selected area (Quito airport) with a high density of cirrus clouds were computed (Fig. 10).

NDVI values were compared to the MODIS MOD13Q1 product and resampled to a spatial resolution of 30 m to enable them to be related to Landsat data. The MODIS product is of a nearer date (07/28/2014) to the Landsat-8 image (Fig. 11(a)). The validation compares the reference NDVI product (MODIS MOD13Q1 resampled) and the NDVI computed through the Landsat-8 image. NDVI values are computed considering the original surface reflectance of the Landsat-8 image (Fig. 11(b)) and the surface reflectance of the images after applying the two algorithms for removing cirrus clouds: i) ACRM (Fig. 11(c)) and ii) ICA (Fig. 11(d)).

In order to compare MODIS NDVI and the other NDVI computations, a linear regression was established to obtain R^2 coefficients, and the results showed that the highest R^2 (0.426) is obtained after applying ACRM. On the other hand, the lowest coefficient is obtained after applying ICA with an R^2 value of 0.262 (Table 6).

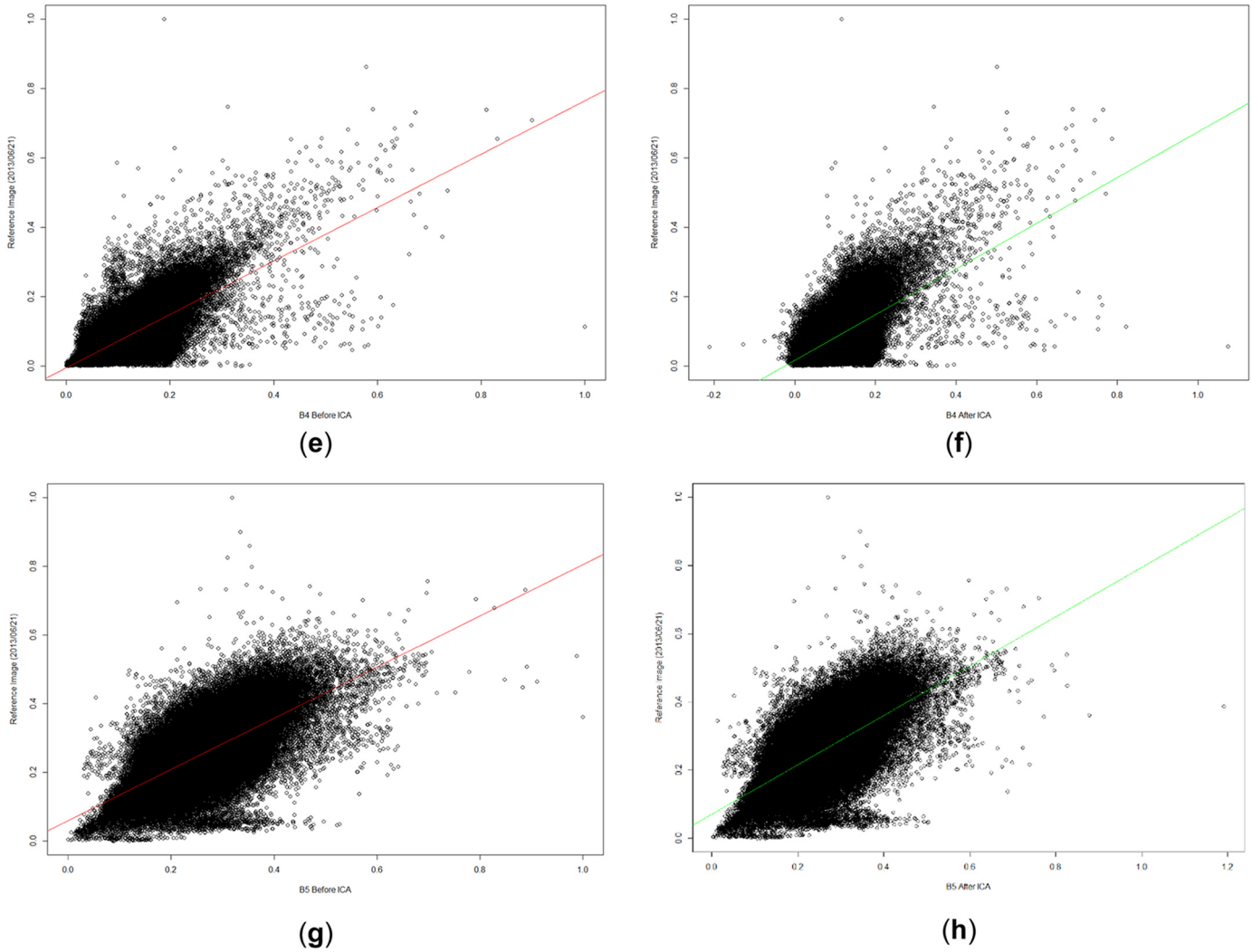


Fig. 9. (*continued*)

Table 5
Linear Regression. R^2 coefficients before and after ICA computation.

Band	R^2 before	R^2 after
B2	0.43	0.20
B3	0.49	0.26
B4	0.53	0.33
B5	0.49	0.47

3.4. Improvement of ACRM

According to the preliminary results (Table 6), the ACRM algorithm yielded the highest R^2 to calculate environmental indices; nevertheless, one improvement of the ACRM method was developed to remove clouds in Landsat-8 OLI images of high-elevation areas (Xu et al., 2014). This development attempts to find the best-fit slope in the ACRM algorithm, established in Eq. (3), to remove clouds in order to compute environmental indices. When ACRM was applied to an image of Quito, the slope parameter presented low values, which led us to conclude that the correction to remove clouds does not work properly when it takes values close to 0 (Table 2).

A previous work used a fixed slope value (Alvarez et al., 2017). The main improvement in the ACRM algorithm was to find the highest R^2 coefficients in the homogeneous zones and the best-fit slope to remove

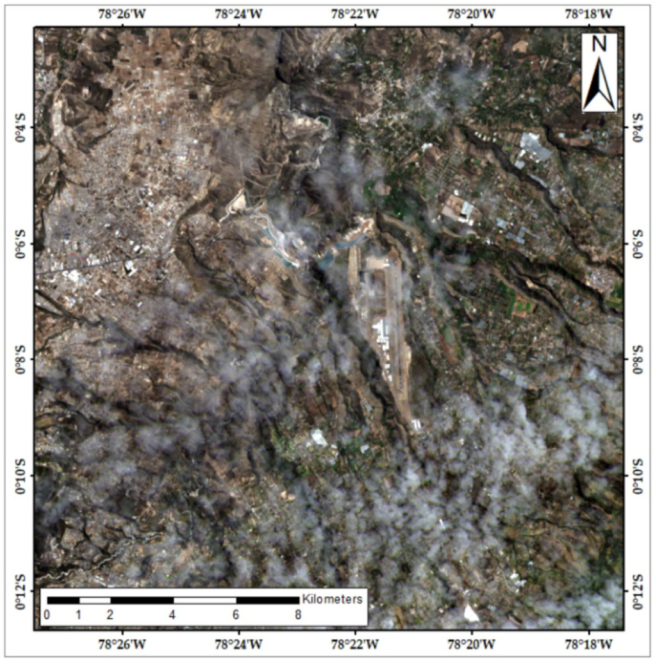


Fig. 10. Area evaluated in Quito airport to compute NDVI (Landsat-8 image from 07/26/2014).

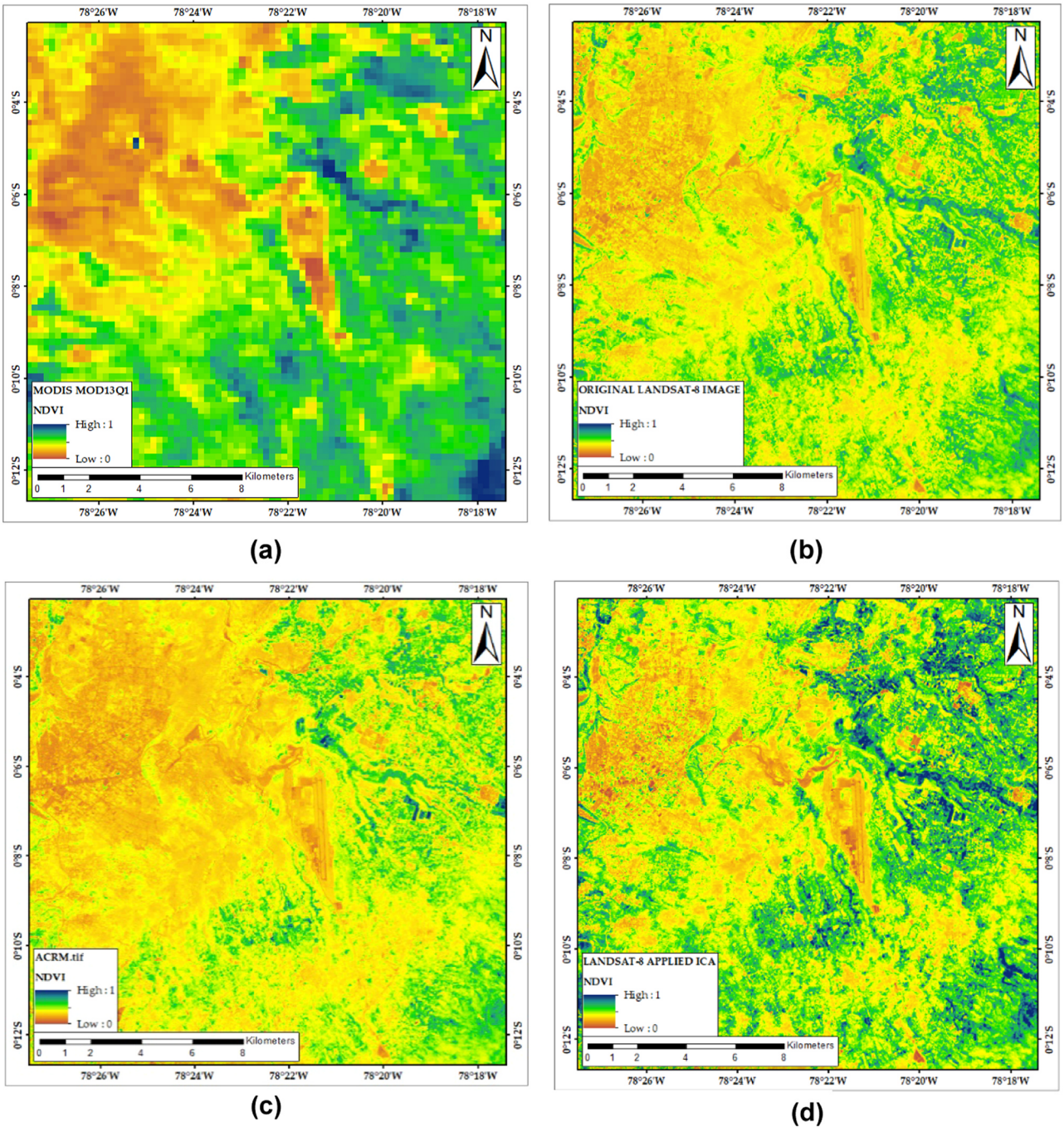


Fig. 11. NDVI computed from **(a)** MODIS NDVI 30 m resampled image; **(b)** original Landsat-8 image; **(c)** Landsat-8 image after cloud correction using the ACRM algorithm; **(d)** Landsat-8 image after cloud correction using the ICA algorithm.

Table 6
Linear Regression between MODIS NDVI and NDVI computed from each cloud removal method.

NDVI Computation with	R^2
Original Image with Surface Reflectance Data	0.396
After ACRM algorithm	0.428
After ICA algorithm	0.262

clouds. Several slope values from 0 to 100 (in increments of 0.1) were tested. Therefore, the improvement was to find the highest R^2 with the fittest slope testing several slopes values. This procedure was implemented in R Studio software.

To compare and validate the best-fit slope, NDVI was computed for the original image (07/26/2014) after applying the ACRM algorithm and compared with the MODIS NDVI, resulting in the highest R^2 (0.5077) with a slope value of 2.9 (Fig. 12).

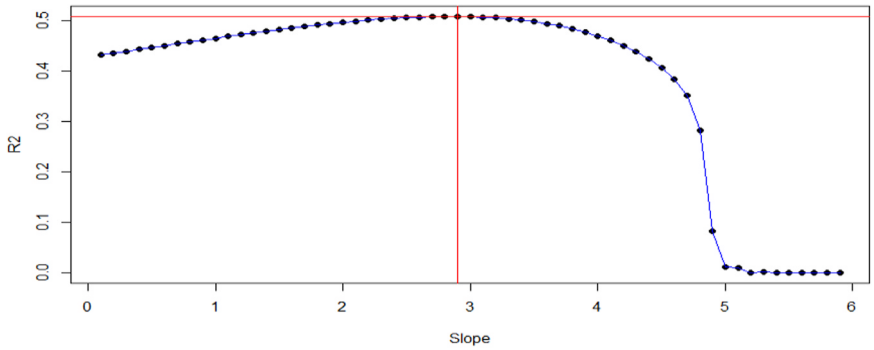


Fig. 12. Comparison between NDVI obtained using ACRM for each slope tested (dots) with the MODIS NDVI. The red lines indicate the highest R^2 and the corresponding slope.

The slope value of 2.9 allowed to a visualization without clouds (Fig. 13 and 14). However, this value is not necessarily the same in each case. The slope value must be investigated for each case, in order to find the best fit to the corresponding area and image.

The results of comparing the R^2 between the different methods are shown in Fig. 15. The improved ACRM shows the highest R^2 value (0.5077), and visually, it removes clouds to yield a clean image (Fig. 13(d)). Thus, the improved ACRM works satisfactorily over the study area.

In order to validate the ACRM, a new image (11/10/2013) with similar properties was used in the same area. The results show a higher R^2 (0.5283) with a slope value of 2.8 in the ACRM (Fig. 16).

4. Discussion and conclusion

Two algorithms, ACRM and ICA, were employed to remove cirrus clouds in Landsat-8 images with the cirrus band (B9) (Department of the Interior U.S. Geological Survey, 2016), in Quito city. The main advantage of these two methods is that they do not use additional images to patch data, in contrast to other methods (Cheng et al., 2014; Lin et al., 2014; Lv et al., 2016; Wu et al., 2016). These methods use the same image to remove thin cloud without the insertion of pixel values from other images. In this work, because cirrus clouds could have a great impact in the computation of environmental indices such as NDVI, these two methods were tested and compared with the aim of evaluating their applicability to accurately compute NDVI for an area located in the Andean region.

ACRM generated satisfactory results for images with conditions similar to Sydney (Xu et al., 2014). The same original image of Sydney was used to reproduce the correct application of ACRM, which yielded an R^2 coefficient higher than 0.95, with slopes higher than 1. These satisfactory results were also evident from visual inspection, because clouds were adequately removed (Fig. 6(d)). When the ACRM algorithm was tested for images of Quito from different dates, the results showed R^2 coefficients higher than 0.90 in most of the cases but with low slope values (lower than 0.1 in most of the cases for all bands) (Table 3). The low slope values indicate poor correction. Moreover, it is evident from visual inspection that this algorithm does not remove the cirrus clouds over the images (Fig. 5). Another area, Pedernales, was chosen to test the algorithm because it has similar characteristics to Sydney. The results for this area are also unsatisfactory for the clouds removal (Fig. 6b).

The other algorithm tested to remove cirrus clouds was ICA (Shen et al., 2015a, 2015b), which is a blind source method that attempts to obtain the cloud component of images (Hyvärinen and Oja, 2000). All ICs contain free pixel data and cloud noise, and the noise should be removed, considering all image data to have a non-Gaussian distribution (Hyvärinen et al., 2001). ICA was tested for images of Quito, and the results were compared with a cloud-free image (image with surface reflectance data). The results are unsatisfactory because the correlation was worse than the case without applying ICA (Table 4). For example, in band 4, the R^2 value obtained in comparison with the cloud-free image was 0.33; the value without applying ICA was 0.53.

In order to validate the results, NDVI was computed. In the first approximation, the results were compared with a reference image product (MODIS MOD13Q1). The results showed the highest R^2 when the ACRM algorithm was applied; these values were higher than those obtained with ICA or those of the surface reflectance data. Finally, an improvement to ACRM was proposed. This algorithm had two main objectives: (i) visually remove clouds and (ii) improve the pixel values to compute environmental indices. The ACRM algorithm was improved, so that the homogeneous area has the highest R^2 coefficient value and the slope should be significant to reduce the density of cirrus clouds. In the case of the study area (Quito), the first condition was achieved with a high R^2 coefficient between Landsat multispectral bands and band 9 in a homogeneous area (Table 1). The challenge was to achieve cloud correction using ACRM. Therefore, we tested different slope values (Alvarez et al., 2017) between 0 and 100, and the best-fit slope value of 2.9 was obtained. This approach proved to be a good alternative to the previous algorithms tested (Fig. 13). In order to validate this new approach, the NDVI values were computed and compared with the reference NDVI values (MODIS). This new approach yielded higher R^2 values (Figs. 15 and 16). The ACRM Improved using the highest R^2 value can approximate to other products ready to use like MODIS NDVI, finding a better relationship than other algorithms or methods, and a considerable best performance, since can be applied to Landsat 8 data, which have a spatial resolution of 30 m.

The preliminary results show that the algorithms to remove cirrus clouds (ACRM and ICA) do not work properly in the geographical conditions considered in this study, leading us to suppose that there are other factors such as altitude and closeness to the equator that influence the results. Therefore, future research should focus on testing these

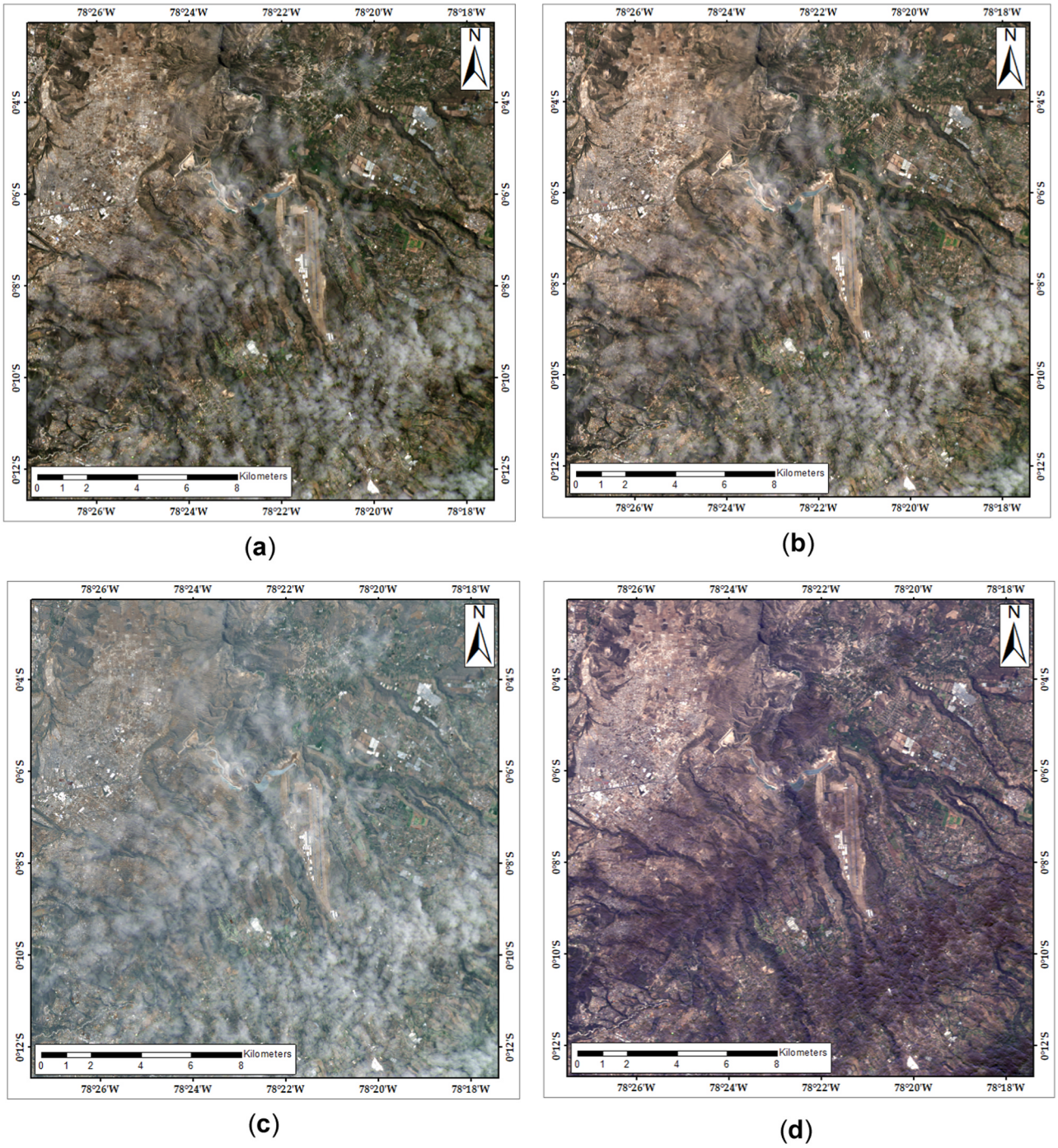


Fig. 13. Images of Quito airport used to compute NDVI (based on Landsat-8 image from 07/26/2014) (**a**) original image; (**b**) image obtained after applying the ACRM algorithm; (**c**) image obtained after applying the ICA algorithm; (**d**) image obtained after applying the improved ACRM algorithm.

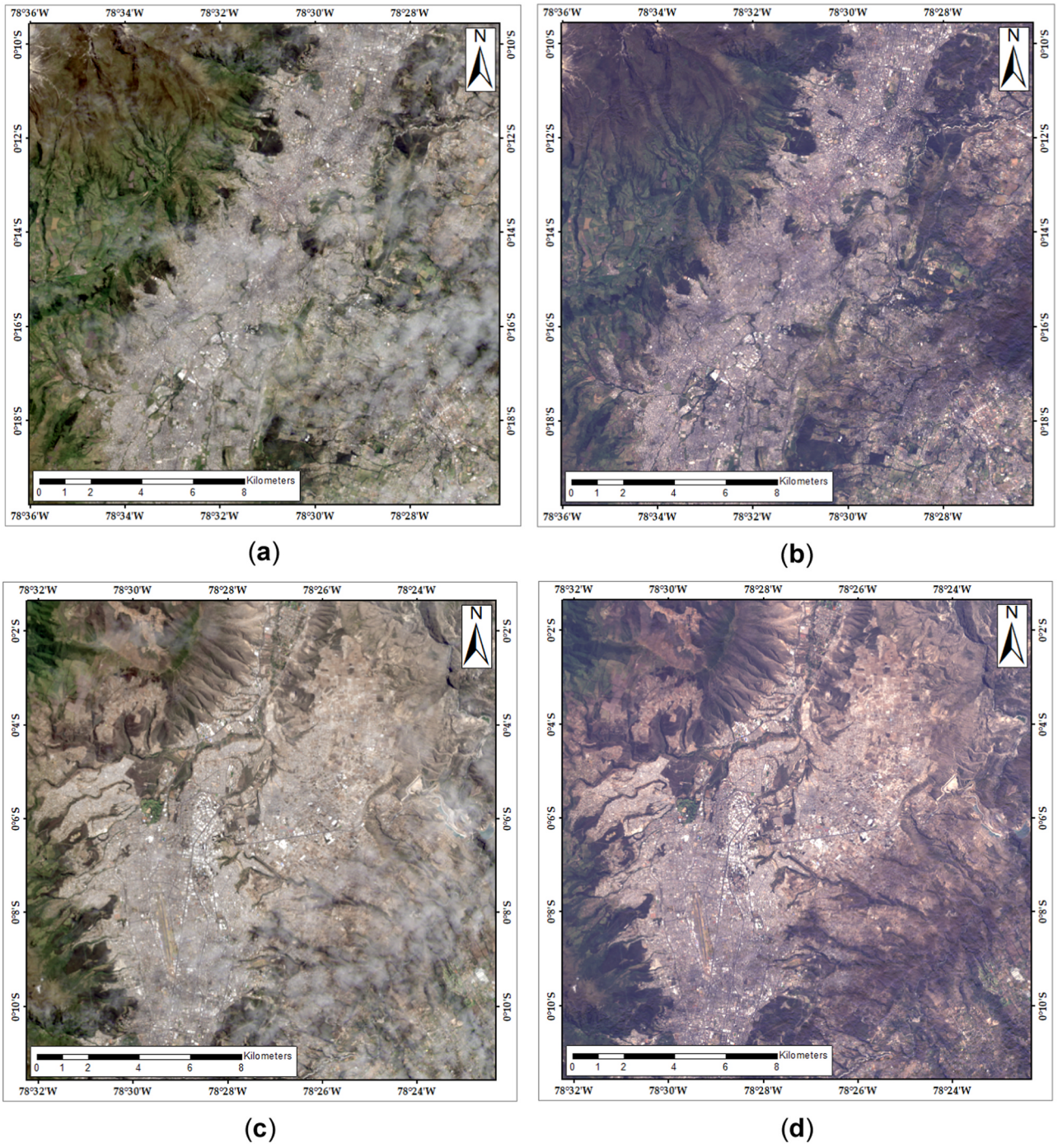


Fig. 14. Comparison of result applying the ACRM improvement (b),(d) in different regions vs. the surface reflectance image (a),(c).

algorithms in different regions around the world to determine the best method for each area or to identify better alternatives to improve the cloud removal algorithms. Moreover, in some parts of the world such as Quito, Landsat images are affected by a high cloud density throughout the year, limiting the time frame to obtain phenology data at a spatial resolution of 30 m. Nevertheless, the ACRM improved can help in a more accurate computation of environmental indexes when compared to other algorithms or methods.

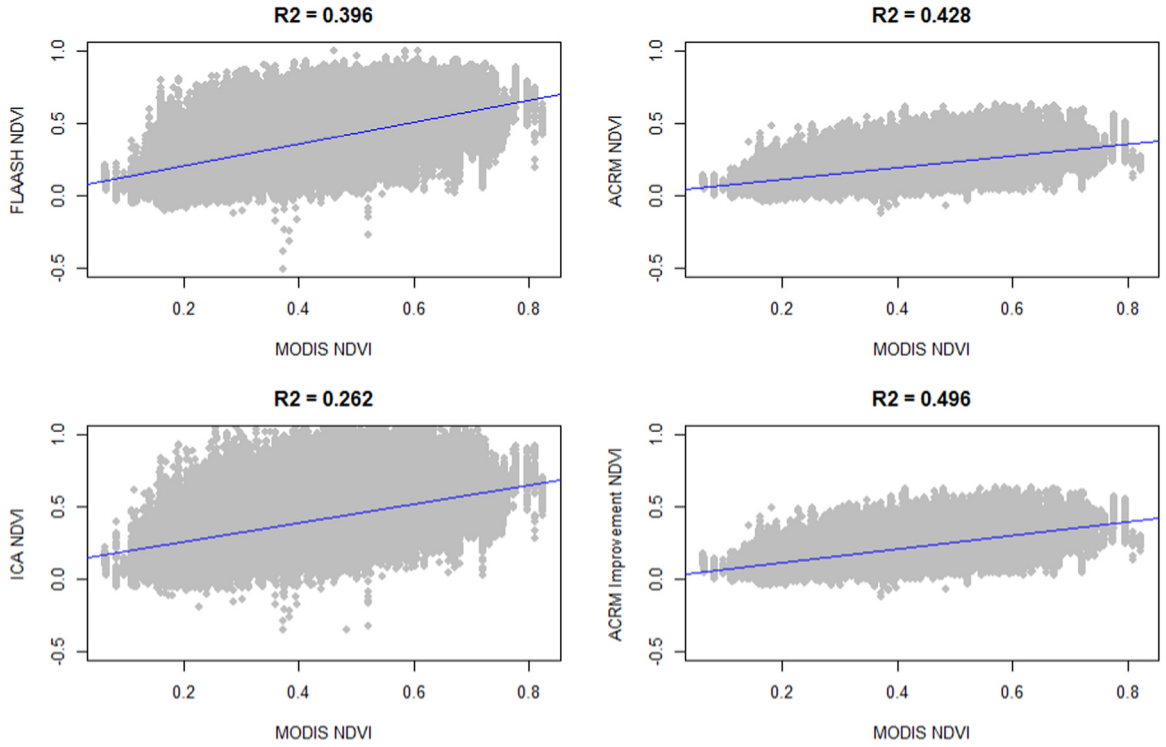


Fig. 15. Comparison between MODIS MOD13Q1 and the different NDVI values obtained from the application of the different algorithms in the Landsat-8 image (07/26/2014) for removing clouds.

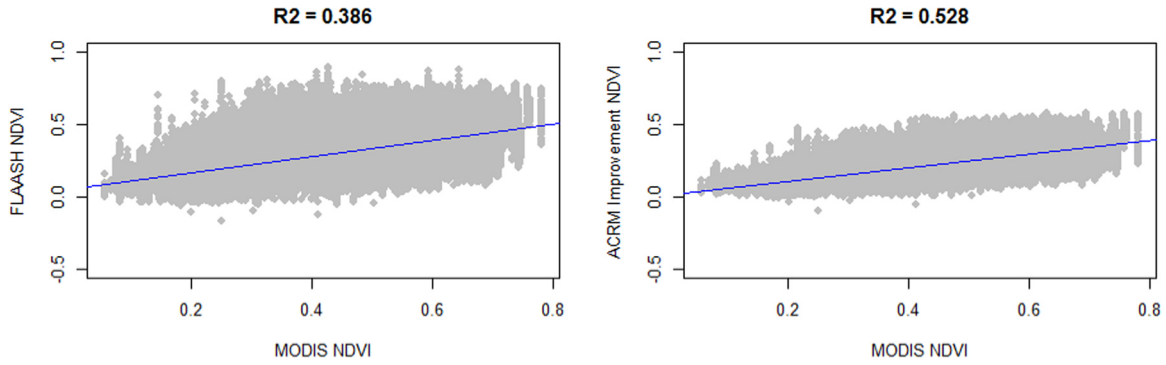


Fig. 16. Comparison between MODIS MOD13Q1 and the NDVI value obtained from Original Surface Reflectance data (FLAASH Correction applied) and ACRM improved in the Landsat-8 image (11/10/2013) for removing cirrus clouds.

Funding

The study is part of a PhD thesis in Surveying Engineering at the University of Porto, Portugal, supported by the Salesian Polytechnic University, Ecuador. This work was supervised at the University of Porto by Prof. Ana Cláudia Teodoro.

Declaration of interest

The authors declare no conflict of interest.

Author contributions

5) Investigation: Cesar I. Alvarez-Mendoza, 10) Supervision: Ana Teodoro, 2) Data curation: Lenin Ramirez-Cando.

References

Allred, C.L., Jeong, L.S, Chetwynd, J.H., 1994. Flaash, a modtran4 atmospheric correction package, 4.

Alvarez, C.I., Teodoro, A., Tierra, A., 2017. Evaluation of automatic cloud removal method for high elevation areas in Landsat 8 OLI images to improve environmental indexes computation SPIE, ed. Proc. SPIE 10428, Earth Resources and Environmental Remote Sensing/GIS Applications VIII 1042809, 10428, pp. 1042809–1042812.

Asner, G.P., 2001. Cloud cover in Landsat observations of the Brazilian Amazon. Int. J. Remote Sens. 22 (18), 3855–3862.

Bivand, R., Keitt, T., Rowlingson, B., 2016. rgdal: Bindings for the Geospatial Data Abstraction Library. Available at: ⟨https://cran.r-project.org/package=rgdal⟩.

Baldock, J.W., 1982. Geology of Ecuador—Explanatory Bulletin of the National Geological Map of the Republic of Ecuador. Ministerio de Recursos Naturales y Energéticos, Dirección General de Geología y Minas, Quito, pp. 70.

Gao, Bo-Cai, Li, Rong-Rong, 2012. Removal of thin cirrus scattering effects for remote sensing of ocean color from space. IEEE Geosci. Remote Sens. Lett. 9 (5), 972–976.

Cheng, Q., et al., 2014. Cloud removal for remotely sensed images by similar pixel replacement guided with a spatio-temporal MRF model. ISPRS J. Photogramm. Remote Sens. 92, 54–68.

Department of the Interior U.S. Geological Survey, 2016. Landsat 8 (L8) Data Users Handbook.

ENVI, 2009. ENVI Atmospheric Correction Module: QUAC and FLAASH user's guide. Module Version, p.44.

Fernández, G., et al., 2015. Land cover change in the Andes of southern Ecuador—patterns and drivers. Remote Sens. 7 (3), 2509–2542.

Gao, B. et al., 1998. Spectral Region Using the Sensitive 1. 375 Cirrus Detecting Channel, 103 (98), pp.169–176.

Gao, B., Li, R., 2017. Removal of thin cirrus scattering effects in landsat 8 OLI images using the cirrus detecting channel. Remote Sens. 9 (8), 834.
Greenberg, J.A., Mattiuzzi, M., 2015. gdalUtils: Wrappers for the Geospatial Data Abstraction Library (GDAL) Utilities. Available at: ⟨https://cran.r-project.org/package=gdalUtils⟩.
Hashim, M., Pour, B.A., Wei, C.K., 2014. Comparison of ETM+ and MODIS data for tropical forest degradation monitoring in the Peninsular Malaysia. Indian Soc. Remote Sens. 42 (2), 383–396.
Weier, J., Herring, D, 2000. Measuring vegetation (NDVI & EVI). Available at: ⟨https://earthobservatory.nasa.gov/Features/MeasuringVegetation/measuring_vegetation_1.php⟩ (Accessed 12 July 2017).
Hijmans, R.J., 2016. raster: Geographic Data Analysis and Modeling. Available at: ⟨https://cran.r-project.org/package=raster⟩.
Huadong, D., Yongqi, W., Yaming, C., 2009. Studies on cloud detection of atmospheric remote sensing image using ICA algorithm. Computer 1–4.
Huete, A., et al., 2002. Overview of the radiometric and biophysical performance of the MODIS vegetation indices. Remote Sens. Environ. 83 (1–2), 195–213.
Hyvärinen, A., Karhunen, J., Oja, E., 2001. Independent component analysis. Appl. Comput. Harmon. Anal. 21 (1), 135–144.
Hyvärinen, A., Oja, E., 2000. Independent component analysis: algorithms and applications. Neural Netw. 13 (4–5), 411–430.
Instituto Nacional de Meteorología e Hidrología, 2016. Boletín Climatológico Anual 2015.
Ji, C.Y., 2008. Haze reduction from the visible bands of LANDSAT TM and ETM+ images over a shallow water reef environment. Remote Sens. Environ. 112 (4), 1773–1783.
Ju, J., Roy, D.P., 2008. The availability of cloud-free Landsat ETM+ data over the conterminous United States and globally. Remote Sens. Environ. 112 (3), 1196–1211.
Kuenzer, C., Dech, S., Wagner, W., 2015. Remote Sensing Time Series Revealing Land Surface Dynamics: status Quo and the Pathway Ahead. Springer, Cham, pp. 1–24.
Lillesand, T., Kiefer, R.W., Chipman, J., 2015. Remote Sensing and Image Interpretation, 7th ed. Wiley (736 p).
Lin, C.-H., et al., 2014. Patch-based information reconstruction of cloud-contaminated multitemporal images. IEEE Trans. Geosci. Remote Sens. 52 (1), 163–174.
Lv, H., Wang, Y., Shen, Y., 2016. An empirical and radiative transfer model based algorithm to remove thin clouds in visible bands. Remote Sens. Environ. 179, 183–195.
Lv, H., Wang, Y., Yang, Y., 2018. Modeling of thin-cloud TOA reflectance using empirical relationships and two landsat-8 visible band data. IEEE Trans. Geosci. Remote Sens. 99, 1–12.
Mandanici, E., et al., 2015. Comparison between empirical and physically based models of atmospheric correction. Proc. SPIE - Int. Soc. Opt. Eng. 9535, 95350E.
Mishra, N.B., Mainali, K.P., 2017. Greening and browning of the Himalaya: spatial patterns and the role of climatic change and human drivers. Sci. Total Environ. 587–588, 326–339.
Pour, B.A., Hashim, M., 2017. Application of Landsat-8 and ALOS-2 data for structural and landslide hazard mapping in Kelantan, Malaysia. Nat. Hazards Earth Syst. Sci. 17 (7), 1285–1303.
R Core Team, 2016. R: A Language and Environment for Statistical Computing. Available at: ⟨http://www.r-project.org/⟩.
Rajitha, K., Prakash Mohan, M.M. & Varma, M.R.R., 2015. Effect of cirrus cloud on normalized difference Vegetation Index (NDVI) and Aerosol Free Vegetation Index (AFRI): A study based on LANDSAT 8 images. ICAPR 2015 - 2015 8th International Conference on Advances in Pattern Recognition, pp. 2–6.
Rees, W.G., 2012. Physical Principles of Remote Sensing. Cambridge University Press.
Richter, R., Wang, X., Bachmann, M., Daniel Schläpfer, D., 2011. Correction of cirrus effects in Sentinel-2 type of imagery. Int. J. Remote Sens. 32 (10), 2931–2941.
Roy, D.P., et al., 2016. Characterization of Landsat-7 to Landsat-8 reflective wavelength and normalized difference vegetation index continuity. Remote Sens. Environ. 185, 57–70.
Shen, H., Li, H., Qian, Y., Zhang, L., Yuan, Q., 2014. An effective thin cloud removal procedure for visible remote sensing images. ISPRS J. Photogramm. Remote Sens. 96, 224–235.
Shen, Y., Wang, Y., Lv, H., Qian, J., 2015a. Removal of thin clouds in landsat-8 OLI data with independent component analysis. Remote Sens. 7 (9), 11481–11500.
Shen, Y., Wang, Y., Lv, H., Li, H., 2015b. Removal of thin clouds using Cirrus and QA bands of Landsat-8. Photogramm. Eng. Remote Sens. 81 (9), 721–731.
Solano, R., et al., 2010. MODIS Vegetation Index User's Guide (MOD13 Series)., 2010(May).
Stephens, G.L., 2005. Cloud feedbacks in the climate system: a critical review. J. Clim. 18 (2), 237–273.
Tucker, C.J., 1979. Red and photographic infrared linear combinations for monitoring vegetation. Remote Sens. Environ. 8 (2), 127–150.
USGS, 2013. User Guide Landsat 8 Operational Land Imager (Oli), (July), pp. 1–16. Available at: ⟨http://landsat.usgs.gov/documents/lldope_toolbelt_userguide_oliqa.pdf⟩.
Wu, M., et al., 2016. An improved high spatial and temporal data fusion approach for combining Landsat and MODIS data to generate daily synthetic Landsat imagery. Inf. Fusion 31, 14–25.
Xu, M., Jia, X., Pickering, M., 2014. Automatic cloud removal for Landsat 8 OLI images using cirrus band. International Geoscience and Remote Sensing Symposium (IGARSS), (September 2016), pp. 2511–2514.
Zambrano, F., et al., 2016. Sixteen years of agricultural drought assessment of the BioBío region in chile using a 250 m resolution vegetation condition index (VCI). Remote Sens. 8 (6), 530.
Zhu, Z., Woodcock, C.E., 2012. Object-based cloud and cloud shadow detection in Landsat imagery. Remote Sens. Environ. 118, 83–94.

environments

MDPI

Article

Assessment of Remote Sensing Data to Model PM10 Estimation in Cities with a Low Number of Air Quality Stations: A Case of Study in Quito, Ecuador

Cesar I. Alvarez-Mendoza [1,2,*], Ana Claudia Teodoro [1,3], Nelly Torres [2] and Valeria Vivanco [2]

[1] Department of Geosciencies, Environment and Land Planning, Faculty of Sciences, University of Porto, Rua Campo Alegre 687, Porto 4169-007, Portugal

[2] Grupo de Investigación Ambiental en el Desarrollo Sustentable GIADES, Carrera de Ingeniería Ambiental, Universidad Politécnica Salesiana, Quito 170702, Ecuador

[3] Earth Sciences Institute (ICT), Pole of the FCUP, University of Porto, Porto 4169-007, Portugal

* Correspondence: calvarezm@ups.edu.ec; Tel.: +593-9-84647745

Received: 14 June 2019; Accepted: 19 July 2019; Published: 21 July 2019

Abstract: The monitoring of air pollutant concentration within cities is crucial for environment management and public health policies in order to promote sustainable cities. In this study, we present an approach to estimate the concentration of particulate matter of less than 10 μm diameter (PM10) using an empirical land use regression (LUR) model and considering different remote sensing data as the input. The study area is Quito, the capital of Ecuador, and the data were collected between 2013 and 2017. The model predictors are the surface reflectance bands (visible and infrared) of Landsat-7 ETM+, Landsat-8 OLI/TIRS, and Aqua-Terra/MODIS sensors and some environmental indexes (normalized difference vegetation index—NDVI; normalized difference soil index—NDSI, soil-adjusted vegetation index—SAVI; normalized difference water index—NDWI; and land surface temperature (LST)). The dependent variable is PM10 ground measurements. Furthermore, this study also aims to compare three different sources of remote sensing data (Landsat-7 ETM+, Landsat-8 OLI, and Aqua-Terra/MODIS) to estimate the PM10 concentration, and three different predictive techniques (stepwise regression, partial least square regression, and artificial neuronal network (ANN)) to build the model. The models obtained are able to estimate PM10 in regions where air data acquisition is limited or even does not exist. The best model is the one built with an ANN, where the coefficient of determination (R^2 = 0.68) is the highest and the root-mean-square error (RMSE = 6.22) is the lowest among all the models. Thus, the selected model allows the generation of PM10 concentration maps from public remote sensing data, constituting an alternative over other techniques to estimate pollutants, especially when few air quality ground stations are available.

Keywords: remote sensing; air quality modeling; air quality monitoring; PM10; LUR

1. Introduction

Due to some factors such as air pollutants permanency over the time, the air quality has decreased in recent years, all over the world. One of the direct indicators of air quality is particulate matter with an aerodynamic diameter lower than 10 μm, usually called PM10 [1]. It is well-known that PM10 has a negative environmental impact on outdoor air quality and that it that is linked to public health problems such as cardiovascular and respiratory diseases [2,3]. Many cities around the world are monitoring PM10 in order to prevent environmental problems. However, this monitoring process needs to be improved in order to establish reliable environmental policies [4]. Thus, understanding the spatial distribution of PM10 requires a scientific and accurate basis to locate the possible sources of this pollutant in cities, in order to avoid environmental problems linked to air quality.

The air quality monitoring network (AQMN) is a classical procedure to monitor PM10 in cities. However, some difficulties are found, for instance, high maintenance cost by station [5], a low quantity of stations in large cities, or non-representative spatial distribution [6]. An alternative could be high resolution air ground measures with the implement of low-cost sensors [7,8], however, they are an investment of the local governments, and most of the time is not possible to realize it. An example of where there is insufficient information provided by AQMN stations and a lack of PM10 measures is in Quito, Ecuador [9–12], where there is not enough information to establish environmental strategies. Quito, the capital of Ecuador, is a special geographic zone, considering its high elevation altitude (2800 m) in the middle of the Andean region. Considering the difficulties of a city like Quito, one valid alternative to complement AQMN monitoring is applying land-use regression models (LUR) [13]. LUR models use different geographical variables as predictors (remote sensing data, meteorological data, road density, vehicular traffic, land use, emission inventory, etc.) [13–16]. However, oftentimes this information cannot be easily accessed. Moreover, these geographical variables are not frequently updated by government institutions. In the case of remote sensing data, the predictors most commonly used in LUR models to retrieve PM10 are aerosol optical depth (AOD) and normalized difference vegetation index (NDVI) from moderate-resolution imaging spectroradiometer (MODIS) products [17–20]. MODIS products have a low spatial resolution that limits their application in medium or small cities [21–23], but they are an efficient alternative to retrieve pollutants in regional (large cities/regions) or national (countries) areas. Consequently, a possible alternative to MODIS products is Landsat data. Nowadays, the operational Landsat satellites are Landsat-7 and Landsat-8 [24,25]. Landsat data have a higher spatial resolution compared with MODIS (30 m instead of 250 m) [23]. Several strategies to retrieve AOD from Landsat data have already been established [24]. Nevertheless, these strategies require AOD ground station data in the study area to have aerosol information in a medium spatial resolution [25,26]. Considering this limitation, other studies suggest that the visible bands of Landsat sensors can be used to invert PM10 [27]. The strategy proposed in this work is useful and effective when the AOD stations are limited.

In order to construct empirical LUR models, some studies have used multiple linear regression (MLR) [26], considering a subset of variables through the stepwise regression (STW) algorithm [28,29]. Nevertheless, the use of MLR cannot analyze the possible multicollinearity between variables, because we have a high correlation between near bands in the spectrum [30]. Moreover, it is well-known that multicollinearity exists between remote sensing variables [31], producing a source of error in MLR empirical models. However, an alternative that allows the computing of more accurate models, avoiding multicollinearity, is to use partial least square (PLS) regression [32–34] or an artificial neuronal network (ANN) [35]. Generally, ANNs give more accurate results in comparison with traditional linear methods, considering the complexity of modeling air pollutants. Some atmospheric studies use a multilayer perceptron (MLP) in the context of ANN in order to obtain a predictor model [26,36].

In Alvarez-Mendoza et al. [12], only remote sensing data were considered to compute the LUR model based in a MLR without a method to select predictors. In this work, three main objectives are proposed: (i) Using only remote sensing data will be used to establish LUR models without any AOD predictor; (ii) making a comparison between three different remote sensing satellite/sensors (MODIS, Landsat-7, and Landsat-8) to retrieve long-term PM10 considering only a selection of predictors and; (iii) comparing the accuracy of different techniques (STW, PLS, and MLP) in the generation of the predictive models. The two last items are the new contributions of this work.

2. Materials and Methods

2.1. Study Area

The study area is the urban zone of Quito, the capital of Ecuador. Quito comprises 45 urban parishes or *parroquias*, distributed between the latitudes 0°30′ S and 0°10′ N and the longitudes 78°10′ W and 78°40′ W (Figure 1). The average elevation is around 2800 meters above sea level. The city is

located in the middle of the Andean Region. The mean minimum and maximum temperatures are approximately 9.0 °C and 25.4 °C, respectively. On the other hand, Quito is a region without four seasons because it is in the tropical area, near to the equatorial line. This area was chosen considering the influence of nine AQMN stations.

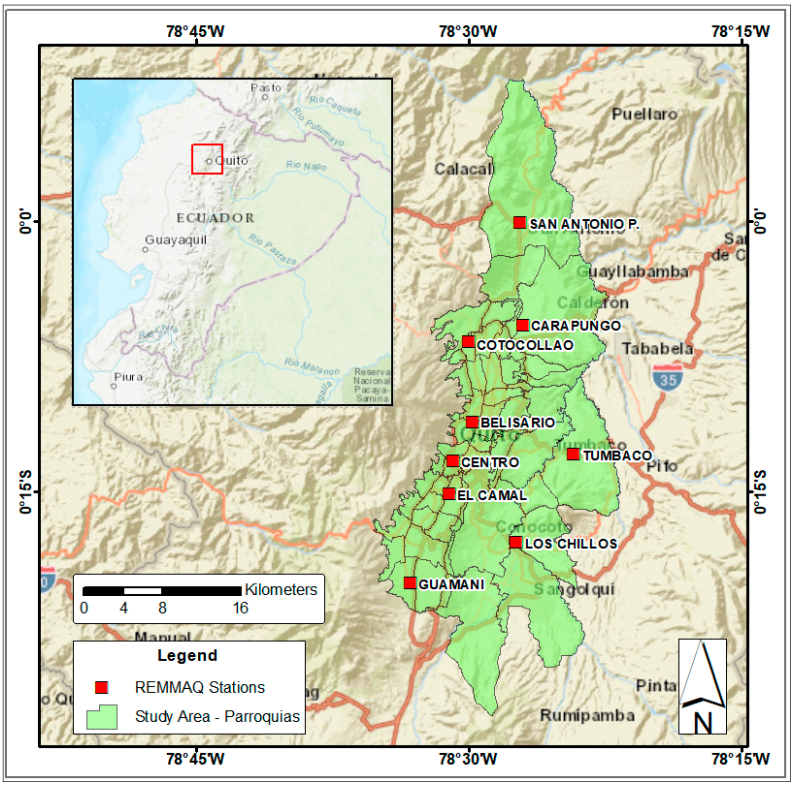


Figure 1. Map of the study area (red dots for REEMAQ (Red Metropolitana de Monitoreo Atmosférico de Quito) stations and green polygons for urban parishes).

2.2. PM10 Data from AQMN Stations

In order to monitor air quality in Quito, nine stations have been acquiring air pollutants since 2002 (Figure 1). Together they form the "Red Metropolitana de Monitoreo Atmosférico de Quito" (REMMAQ) [37]. REEMAQ is the AQMN of Quito, where one of the air pollutants daily measured is PM10. These data are public and free to download (http://www.quitoambiente.gob.ec/ambiente/index.php/datos-horarios-historicos). The PM10 concentration is measured in micrograms per cubic meter ($\mu g/m^3$). In this study, we use three-month-averages from 2013 to 2017, matching with the dates of the remote sensing data (time when the satellite passes over the study area). The main reasons to use three-month-averages are the few available remote sensing data and REMMAQ stations (stations without data in some months or with negative data values). In this study, PM10 three-month-averages are used as the dependent variable.

2.3. Remote Sensing Data Predictors

In this study, three different types of remote sensing data were used to retrieve PM10 between 2013 and 2017: Landsat-7 ETM+, Landsat-8 OLI/TIRS and MODIS/Terra and Aqua (Table 1). The remote sensing data are free to download from the United States Geological Survey (USGS) website (http://earthexplorer.usgs.gov). Moreover, only images with less than 10% cloud cover were considered in the study, because one of the main problems in these regions is the presence of a high cloud density [38,39]. According to this limitation, just 40% of remote sensing data was considered.

Table 1. Characteristics of satellites and sensors used in the study.

Satellite	Sensor	Overpass Time of Satellite	Spatial Resolution
Landsat-7	Enhanced Thematic Mapper Plus (ETM+)	16 days	30 m
Landsat-8	Operational Land Imager (OLI) Thermal Infrared Sensor (TIRS)	16 days	30 m
Terra (EOS AM-1) Aqua (EOS PM-1)	Moderate Resolution Imaging Spectroradiometer (MODIS) MCD43A4	1 to 2 days	500 m

The predictors or independent variables (surface reflectance bands and environmental indexes) are listed in Table 1. The selection of remote sensing predictors was related to their possible correlation with the PM10 concentration [9,40–42]. In the case of the environmental indexes, the most popular indexes in LUR studies to retrieve PM10 were used. They were computed as (1), (2), (3), (4), and (5) in Table 2, respectively.

Table 2. Remote sensing predictors used to build the model for each sensor.

Predictors	Landsat-7	Landsat-8	MODIS
Blue band (B) Green band (G) Red band (R) Near Infrared (NIR) Short Wave infrared (SWIR)	Landsat surface data Level-2	Landsat surface data Level-2	MODIS MOD09A1 MYD09A1 products
Normalized Difference Vegetation Index (NDVI)	$NDVI = \frac{NIR-R}{NIR+R}$ (1)		MODIS MOD13Q1 MYD13Q1products
Normalized Difference Soil Index (NDSI)	$NDSI = \frac{SWIR-NIR}{SWIR+NIR}$ (2)		
Soil-Adjusted Vegetation Index (SAVI)	$SAVI = (1+L)\frac{NIR-R}{NIR+R+L}$ (3) where L represents a minimal change in the soil brightness with a value of 0.5 [43]		
Normalized Difference Water Index (NDWI)	$NDWI = \frac{G-NIR}{G+NIR}$ (4)		
Land Surface Temperature (LST)	$LST = \frac{BT}{\left(1+\left(\frac{\lambda * BT}{\rho}\right) ln\varepsilon\right)} - 273.15$ (5) where BT is the brightness temperature, λ is the center wavelength (Landsat-7 = 11.45 μm, Landsat-8 = 10.8 μm) [44], ρ is a constant and ε is the emissivity [45,46].		MODIS MOD11A1 MYD11A1 products

2.4. LUR Models

LUR models are an alternative to predict the spatialization of air pollutants, particularly when the number of AQMN stations is limited. They use different geographical variables such as roads, traffic information, meteorological and remote sensing data, and other environmental variables, in order to build a model to retrieve air pollutants. However, often several geographical variables are not available. Thus, we should use simple alternatives, such as free remote sensing data, as variables to approach a LUR model.

In most cases, LUR uses MLR to establish the model [47,48]. MLR allows an easy and simple model construction. In our case, the dependent variable is the quarterly PM10 value and the independent variables or spatial predictors are the remote sensing data in each coordinate of the AQMN station, considering the free cloud pixel value. Equation (6) shows the original LUR model, considering all the remote sensing predictors in MLR.

$$PM10 = I + aNDVI - bNDSI - cSAVI + dNDWI - eLST - fB - gG + hR + iNIR + jSWIR + kY - lS \quad (6)$$

where I is the intercept, NDVI is normalized difference vegetation index, NDSI is the normalized difference soil index, SAVI is the soil-adjusted vegetation index, NDWI is the normalized difference water index, LST is the land surface temperature, B is the blue band, G is the green band, R is the red band, NIR is the near infrared band, SWIR is the shortwave infrared band, Y is the year of image acquisition, S is the three-month-averages of image acquisition (January–March—1, April–June—2, July–September—3, and October–November—4), a, b, . . . , l, are the coefficients in each predictor. The other variables are described in Table 1.

Nevertheless, considering that multicollinearity exists between remote sensing variables [31], different predictor techniques should be employed to compute the LUR model. We compare three techniques, namely, MLR with STW, PLS, and ANN, in order to find the fittest model (Figure 2).

In the first model, we use MLR considering an STW. It contemplates different parameters in order to identify the most adequate/influencing variables as predictors. The parameters used to subset the variables are: (i) The residual sum of squares for each model (RSS); (ii) the adjusted regression coefficient R^2 (Adj. R2); (iii) Mallows' Cp (CP) and; (iv) Bayesian information criterion (BIC).

The second model uses PLS with the STW criteria to select the predictors. The main challenge when using PLS is to avoid multicollinearity, finding an alternative when we have few data and a significant number of predictors [49]. PLS generates new latent variables or components in a lineal way.

Finally, the last model uses an ANN in an MLP, with a hidden layer and six hidden nodes to compute the predictive model. The nodes are computed according to [50]. In this model, we use all the predictors. This method is used when the model is complex, giving a different weight to each predictor corresponding to its importance. Additionally, we use a non-linear activation function with backpropagation. The training data to build the MLP consider 75% of the dataset and the remaining 25% for test. We use a backpropagation approach to train the algorithm. The R studio software was used in this study to extract the data and to compute all the models.

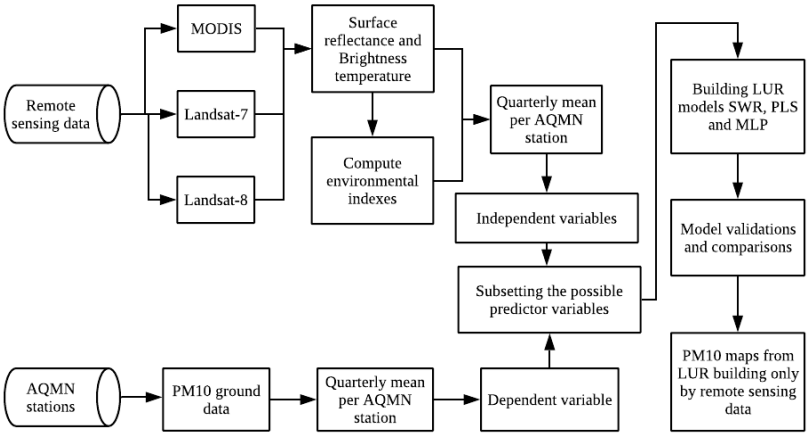


Figure 2. Workflow of the methodology proposed to establish the land use regression (LUR) models.

3. Results

PM10 ground measurements and remote sensing data are matched in a table with the same date. Thus, the unique condition is to consider remote sensing data with less than 10% cloud density. So, the three-month-averages matching tables for each sensor contain 35 observations for Landsat-7, 93 observations for Landsat-8, and 108 observations for MODIS. The main reasons to have only these numbers of observations are the high cloud density in the study area and the incomplete/not available air pollution data. Furthermore, the criteria to select predict variables consider five dependent variables for Landsat-7, eight dependent variables for Landast-8, and six dependent variables for MODIS, for each STW and PLS model, as shown in Table 3. They were obtained according to STW criteria (RSS, Adj. R2, CP, and BIC). The variables common to all the three cases considered are blue band, near infrared (NIR) band, and normalized difference vegetation index (NDVI).

Table 3. Number of observations and predictors per satellite to build the LUR models.

Variable	Landsat-7	Landsat-8	MODIS
No. Observations	35	93	108
No. Predictors	5	8	6
Predictors	NDVI B R NIR S	NDVI SAVI LST B G R NIR Y	NDVI B G R NIR S

The LUR models are computed considering STW and PLS regressions in a linear way and MLP in a non-linear way. They are shown and compared in Table 4 (Equations (7)–(12)). In the case of Landsat-7, the STW shows a coefficient of determination (R^2) of 0.37, the PLS a R^2 of 0.36, and, for MLP, a R^2 of 0.46. The lowest root-mean-square error (RMSE) was obtained for STW with a value of 9.47. For Landsat-8, in STW a R^2 of 0.42 was obtained, and a R^2 of 0.43 for PLS, and a R^2 of 0.68 for MLP (Figure 3). The lowest RMSE obtained was for MLP. Finally, for MODIS, a R^2 of 0.15 for STW, a R^2 of 0.19 for PLS and a R^2 of 0.25 for MLP were obtained. The lowest RMSE was for STW.

Table 4. LUR models for each sensor with different regression techniques. In the case of multilayer perceptron (MLP), the model is not linear.

Sensor	Model	Equation/Method	Coefficient of Determination (R^2)	Root-Mean-Square Error (RMSE)
Landsat-7 ETM+	Stepwise regression (STW)	PM10 = −26.770 + 205.289NDVI − 0.073B + 0.144R − 0.048NIR + 2.270S (7)	0.37	9.47
	Partial least square regression (PLS)	PM10 = 24.786 − 54.369NDVI − 0.059B + 0.049R − 0.008NIR + 2.165S (8)	0.36	10.14
	Multilayer perceptron (MLP)	Non-linear. One hidden layer and six hidden nodes.	0.46	12.69
Landsat-8 OLI/TIRS	STW	PM10 = −4125.506 + 350.130NDVI − 200.334SAVI0.936LST − 0.035B − 0.036G + 0.099R − 0.013NIR + 2.061Y (9)	0.42	9.19
	PLS	PM10 = −4146.508 + 115.816NDVI − 40.465SAVI1.020LST − 0.036B − 0.038G + 0.104R − 0.016NIR + 2.073Y (10)	0.43	9.46
	MLP	Non-linear. One hidden layer and six hidden nodes.	0.68	6.22
MODIS	STW	PM10 = 1.248 + 93.411NDVI + 0.056B − 0.070G + 0.056R − 0.017NIR + 3.190S (11)	0.15	12.91
	PLS	PM10 = 5.661 + 79.106NDVI + 0.060B − 0.072G + 0.050R − 0.014NIR + 3.308S (12)	0.19	12.93
	MLP	Non-linear. One hidden layer and six hidden nodes.	0.25	16.38

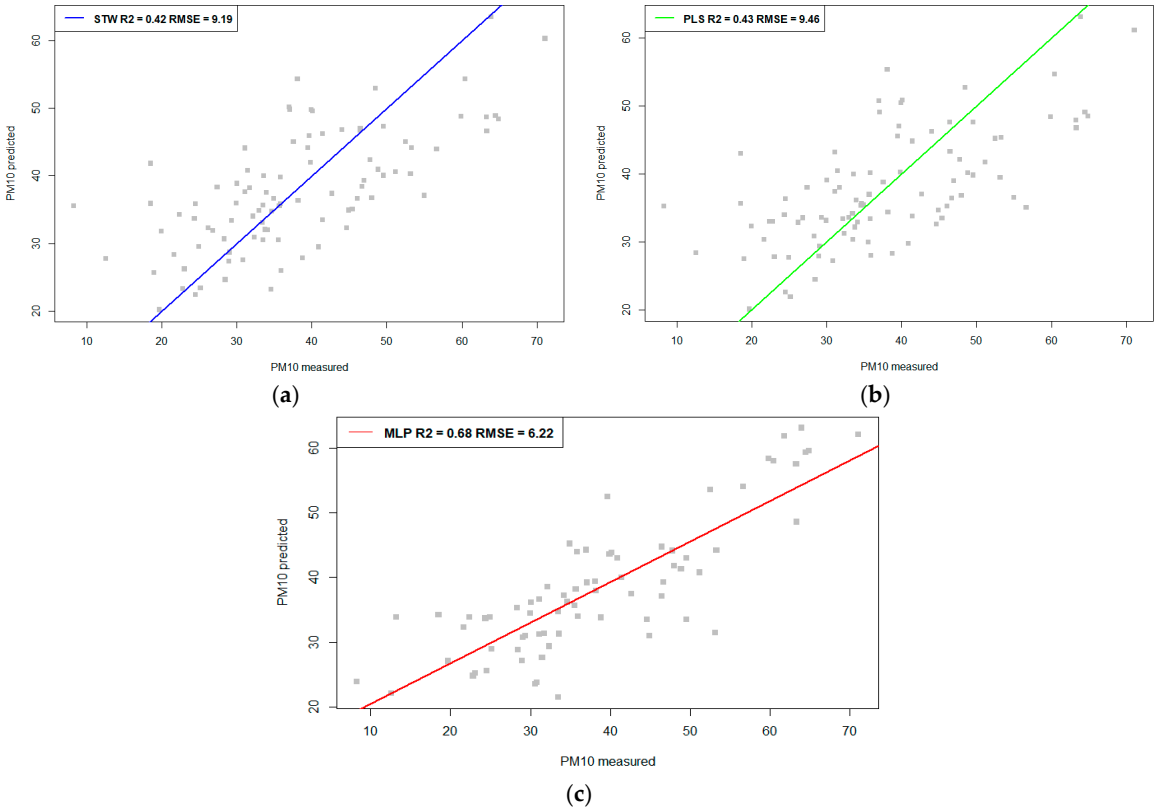


Figure 3. Comparison between R^2 and root mean square error (RMSE) in the model results for Landsat-8 data: (**a**) stepwise regression (STW); (**b**) partial least square (PLS); (**c**) MLP.

The results in Table 4 show that Landsat-8 data with MLP are the fittest model. The MLP employed (Figure 4) has one hidden layer with six hidden nodes.

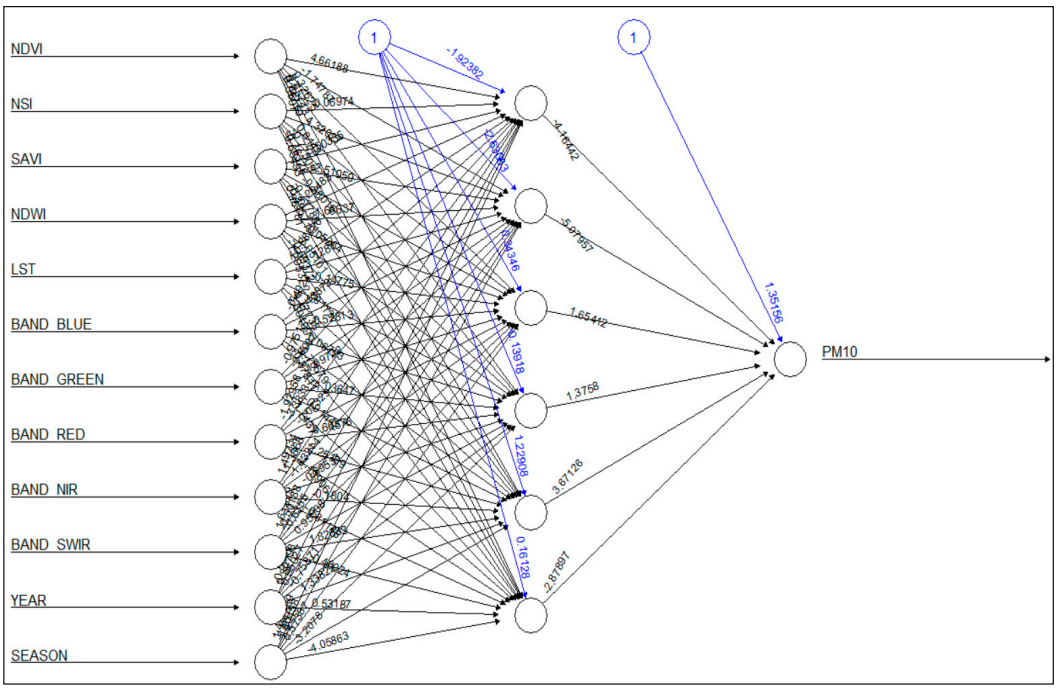


Figure 4. MLP diagram for Landsat-8 data.

Figure 5 shows the relative variable importance according to the assigned weights, where the red band is the most significant in the model, while LST presented the lowest significance.

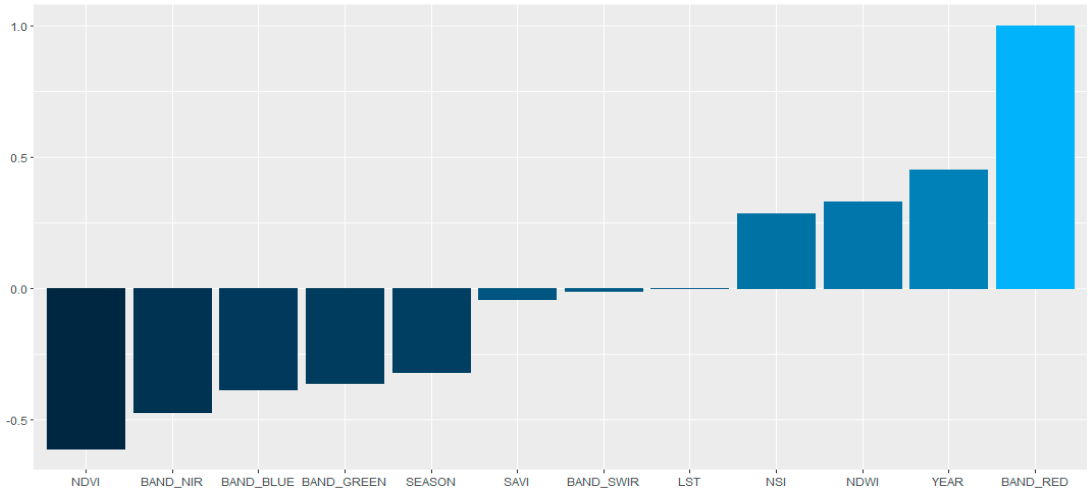
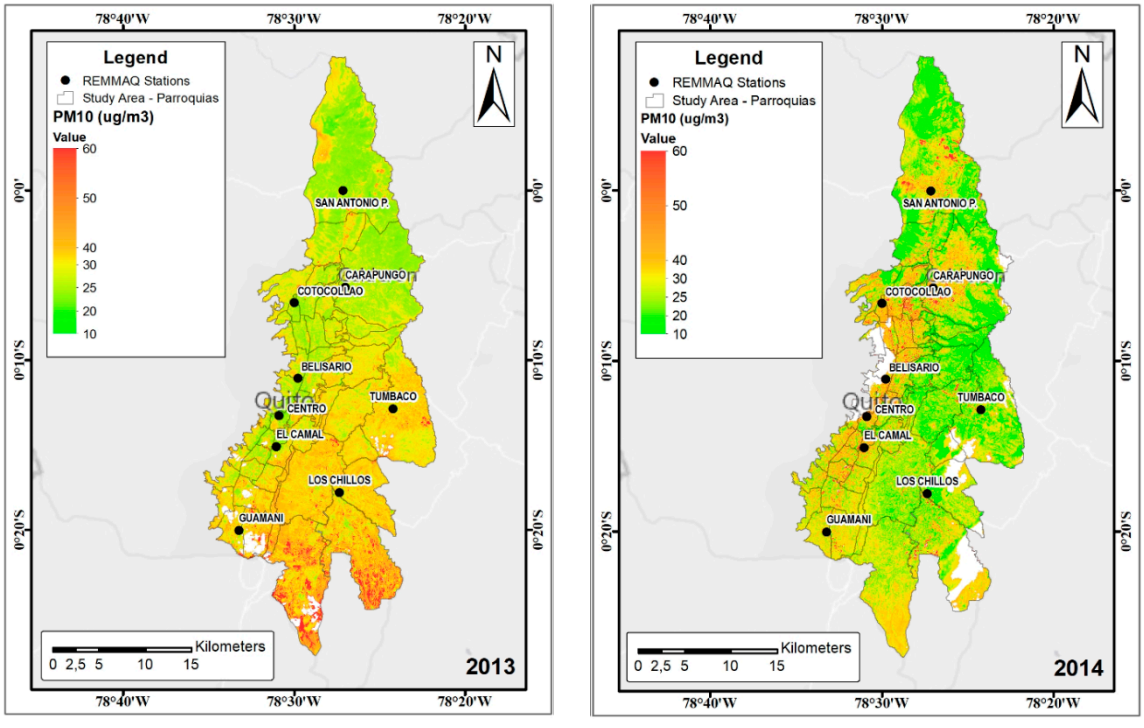


Figure 5. Relative variable importance in Landsat-8 MLP. The scale is between −0.5 and 1, where 0 is the lowest (null) importance.

The Landsat-8 LUR-MLP model is chosen to predict PM10, considering the highest R^2 and the lowest RMSE. In Figure 6, the quarterly maps show the PM10 spatial concentration during 2015, in a color scale in µg/m3. The white gaps showed in the maps are clouds with a high density.



(**a**) (**b**)

Figure 6. *Cont.*

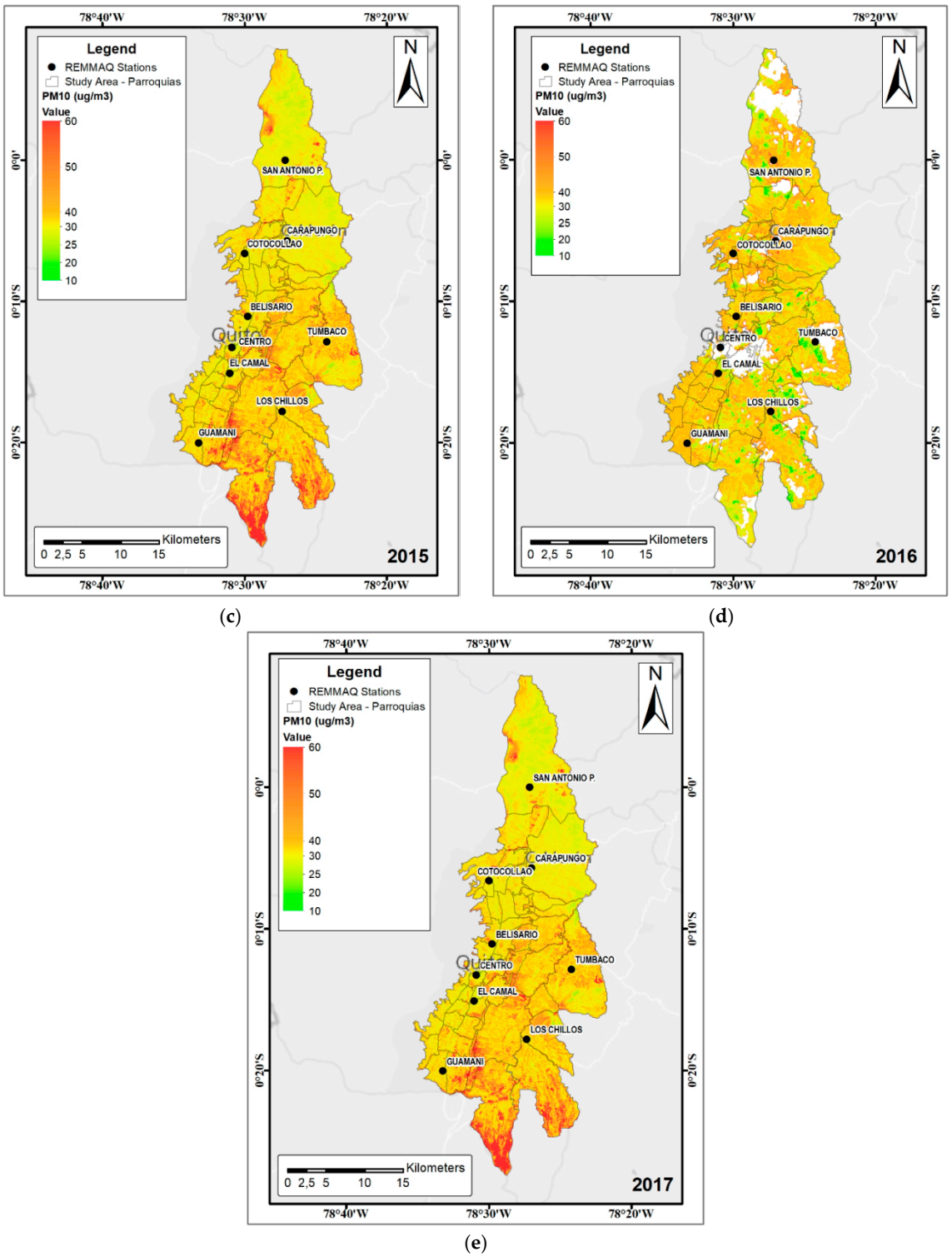


Figure 6. PM10 concentrations during the season 4 (July to September) with Landsat-8 LUR-MLP model in: (**a**) 2013; (**b**) 2014; (**c**) 2015; (**d**) 2016; (**e**) 2017. The white gaps represent areas with a high cloud density.

4. Discussion

As demonstrated in this study, LUR models are an interesting alternative to model air quality, specifically PM10 concentrations, when the in-situ air quality measures are insufficient. Usually, most of the predictors are geographical variables (such as roads), traffic, meteorological data, and others [13].

LUR models are usually applied in small cities or regions where AQMN stations are limited [51], and where spatial interpolation techniques, such as ordinary kriging or inverse distance weighting, cannot be applied, considering the low number of ground measurements available [52]. One of the main problems with these geographic variables is the low accessibility to the data and the time of acquisition. Sometimes, these variables are obsolete, and they are not enough to establish a possible trend.

In this study, we propose an alternative, considering only free remote sensing variables. We apply this approach to the city of Quito, Ecuador, during the period between 2013 and 2017, in order to compare three different satellite data. Quito is growing in new poles. When REEMAQ was established in 2002, Quito did not have its current size and configuration. Now, REEMAQ is an obsolete air quality network, especially in the distribution of stations, which urgently needs improvement. Air pollutant spatial models are techniques based on interpolation or geostatistics approaches, which can be useful if a reasonable number of stations are available with a good spatial distribution [53]. In this study, only nine stations are available. Moreover, in some cases, the data are incomplete during some months. Additionally, according to some authors [7,8], it is possible to have more air ground data with low-cost sensors, however they must be implemented in the cities in order to monitor the air quality. The alternative to improve the air quality model in Quito is to establish different spatiotemporal LUR models, considering only remote sensing data as predictor variables. A preliminary study shows the use of only remote sensing variables but using an MLR in order to build the model. The limitation is the use of all remote sensing predictors without considering the collinearity [12]. In order to establish the models, three different remote sensing data were tested (Landsat-7, Landsat-8, and MODIS) and three techniques for modeling (STW, PLS, and MLP) were employed. The selected variables to compute the model are the visible NIR and SWIR bands of the three sensors, different environmental indices (NDVI, NDSI, SAVI, NDWI), and LST, computed from the data retrieved from each sensor. Most of the studies published use aerosol optical thickness (AOT) derived from MODIS (MOD04) [54] as the input in LUR models, however, this product has a low spatial resolution (3×3 km) [55]. This resolution is not practicable when considering cities like Quito, where the maximum width is near to 10 km. On the other hand, some MODIS products do not have a suitable quality for local studies [56]. Other studies use Landsat-8 combined with AOT ground stations to spatially model the AOT [24]. This could be a good alternative, however in our study area, we do not have access to this information between 2013 and 2017.

Comparing the LUR models established, we found that Landsat-8 is the most adequate sensor to model PM10 concentration, considering the 93 records and according to a previous study [12]. MLP is the fittest alternative to model PM10, with a R^2 of 0.68 and a RMSE of 6.22. In this context, the non-linear model (MLP) has a fitter result when compared to the linear models (STW and PLS) [26]. Therefore, the LUR-MLP model was chosen to map the spatial concentration of PM10 in Quito, between 2013 to 2017. MODIS presents the lowest R^2 with a value of 0.19, considering the PLS regression. This could be related to the lowest spatial resolution. Thus, most of the LUR models use MLR or STW. MLR is easy to implement. However, one of the main problems could be the multicollinearity, because MLR does not analyze the correlation between predictors [57]. On the other hand, the linear PLS helps to avoid the multicollinearity creating new latent variables with few observations [34]. In a future work, a possible combination between STW (in order to select the predictor variables), non-linear PLS (in order to avoid the multicollinearity between remote sensing data), and a machine learning technique (as ANN) can improve the LUR models [58].

In the case of the predictors, all the models present, in all the cases, the variables blue band, NIR, and NDVI. In the case of NDVI, a possible reason is the direct influence of vegetation on the PM10 concentration and distribution [18]. On the other hand, the red band has the most importance in MLP, because there could be a relationship between the retrieval of PM10 with the blue and red bands [27]. In most of the LUR studies, the authors use traffic, roads, meteorological, land use, population, and other predictors, reporting values of R^2 according to the reality of each local [26]. These models also considered different time periods (monthly, quarterly, yearly). The main difference of our approach

is the use of remote sensing data only as predictors, which can replace the necessity to have all geographical variables. Another advantage is the data availability and continuity in order to recompute the LUR models. One of the main limitations of our model is the high cloud density presented in the images during all the year [38], making it complicated to use more data in order to improve the model. However, a future work will intend to have more satellite sensors or to find new alternatives to recover remote sensing data contaminated with clouds [39].

Figure 6 shows variations year by year according to PM10 mean concentration based on in-situ data (REEMAQ Stations). We choose the third season to show the variation year by year (2013–2017), because we have more remote sensing data available (without a high cloud density) during this time-window. According to the results presented in Figure 6, an increasing of PM10 concentration between 2013 to 2017 is notorious in the most of the urban parishes [59]. However, some areas showed a decreasing tendency in some years. The lowest PM10 concentration was found in some peripheral parishes during the 2014 year, because the air stations that influence these parishes (Tumbaco and Los Chillos) had a variation in the concentrations. Thus, Tumbaco and Los Chillos stations are in the east part of the study area and began to present the lower values in 2014 followed by 2013, according to the in-situ measures. After 2014, the PM10 values for these stations began to increase. The main reason could be related to the new operation of the new airport of Quito (2013), and the construction of important road infrastructures around it (end of 2014). Another possible explication is the traffic influence during the last years, particularly in the peripheral areas where an increase was registered since 2015 and also the increase of the population in these areas [60]. In the northern parishes, the stations of San Antonio P. and Carapungo are influenced by the presence of stone and sandy point quarries [61]. The stations Centro, Belisario, and El Camal are in the city downtown, and it is the main reason that an increase of PM10 concentration during the last years is verified in the center parishes.

According to our results, several areas presented concentrations higher than 50 $\mu g/m^3$ (Figure 6), while the World Health Organization (WHO) recommends, in its guidelines, maximum values of 20 $\mu g/m^3$ as an annual mean and 50 $\mu g/m^3$ as a 24-h mean [1]. However, some areas do not show values, due to the high cloud density (white areas in Figure 6). Thus, the PM10 concentration maps from the Landsat-8 LUR-MLP model can help local government decision makers to manage air quality concentration and to organize new policies, specifically in the places where the highest concentrations were identified.

5. Conclusions

In this study, three different satellite datasets were compared to retrieve models of PM10 through LUR, in Quito, Ecuador between 2013 and 2017. Additionally, three techniques were compared to compute the LUR models (SWR, PLS, and MLP). From this work, several conclusions could be taken: (i) It is possible to build empirical models established using only remote sensing variables as predictors without any other geographic variables, as traditional LUR models; (ii) in the case of Quito, the study results show that Landsat-8 provides the most suitable satellite data to retrieve PM10, in comparison with Landsat-7 and MODIS; (iii) MLP allows the obtainment of the most robust result in comparison with the other modeling techniques. MLP is the fittest alternative to model PM10, with a R^2 of 0.68 and a RMSE of 6.22, and; (iv) the MLP model established helps in the spatial mapping of PM10, where in the time window of this study, were found areas with PM10 values higher than the limit established by WHO. Thus, these models are useful in the management of air quality in the city of Quito and can be applied to other locations with similar characteristics.

Author Contributions: Conceptualization, C.I.A.-M.; data curation, N.T. and V.V.; investigation, C.I.A.-M.; supervision, A.C.T.; writing—review and editing, C.I.A.-M.

Acknowledgments: The study is part of a PhD thesis in Surveying Engineering at the University of Porto, Portugal, supported by the Salesian Polytechnic University, Ecuador. This work was supervised at the University of Porto by Prof. Ana Cláudia Teodoro.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. WHO Ambient (Outdoor) Air Quality and Health. Available online: http://www.who.int/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health (accessed on 30 August 2018).
2. Kutlar Joss, M.; Eeftens, M.; Gintowt, E.; Kappeler, R.; Künzli, N. Time to harmonize national ambient air quality standards. *Int. J. Public Health* **2017**, *62*, 453–462. [CrossRef] [PubMed]
3. Kobza, J.; Geremek, M.; Dul, L. Characteristics of air quality and sources affecting high levels of PM10 and PM2.5 in Poland, Upper Silesia urban area. *Environ. Monit. Assess.* **2018**, *190*, 515. [CrossRef] [PubMed]
4. World Health Organization Regional Office for Europe. *Health Effects of Particulate Matter*; World Health Organization Regional Office for Europe: København, Denmark, 2013.
5. Ielpo, P.; Paolillo, V.; de Gennaro, G.; Dambruoso, P.R. PM10 and gaseous pollutants trends from air quality monitoring networks in Bari province: Principal component analysis and absolute principal component scores on a two years and half data set. *Chem. Cent. J.* **2014**, *8*, 14. [CrossRef] [PubMed]
6. Pope, R.; Wu, J. A multi-objective assessment of an air quality monitoring network using environmental, economic, and social indicators and GIS-based models. *J. Air Waste Manag. Assoc.* **2014**, *64*, 721–737. [CrossRef] [PubMed]
7. Capezzuto, L.; Abbamonte, L.; De Vito, S.; Massera, E.; Formisano, F.; Fattoruso, G.; Di Francia, G.; Buonanno, A. A maker friendly mobile and social sensing approach to urban air quality monitoring. In Proceedings of the IEEE SENSORS 2014, Valencia, Spain, 2–5 November 2014; pp. 12–16.
8. Hasenfratz, D.; Saukh, O.; Walser, C.; Hueglin, C.; Fierz, M.; Thiele, L. Pushing the spatio-temporal resolution limit of urban air pollution maps. In Proceedings of the 2014 IEEE International Conference on Pervasive Computing and Communications (PerCom), Budapest, Hungary, 24–28 March 2014; pp. 69–77.
9. Alvarez, C.I.; Padilla Almeida, O.; Álvarez Mendoza, C.I.; Padilla Almeida, O. Estimación de la contaminación del aire por PM10 en Quito a través de índices ambientales con imágenes LANDSAT ETM+. *Rev. Cart* **2016**, 135–147.
10. Cevallos, V.M.; Díaz, V.; Sirois, C.M. Particulate matter air pollution from the city of Quito, Ecuador, activates inflammatory signaling pathways in vitro. *Innate Immun.* **2017**, *23*, 392–400. [CrossRef] [PubMed]
11. Raysoni, A.U.; Armijos, R.X.; Weigel, M.M.; Montoya, T.; Eschanique, P.; Racines, M.; Li, W.-W. Assessment of indoor and outdoor PM species at schools and residences in a high-altitude Ecuadorian urban center. *Environ. Pollut.* **2016**, *214*, 668–679. [CrossRef]
12. Alvarez-Mendoza, C.I.; Teodoro, A.; Torres, N.; Vivanco, V.; Ramirez-Cando, L. Comparison of satellite remote sensing data in the retrieve of PM10 air pollutant over Quito, Ecuador. In Proceedings of the SPIE - The International Society for Optical Engineering, Berlin, Germany, 9 October 2018.
13. Yang, X.; Zheng, Y.; Geng, G.; Liu, H.; Man, H.; Lv, Z.; He, K.; de Hoogh, K. Development of PM2.5 and NO2 models in a LUR framework incorporating satellite remote sensing and air quality model data in Pearl River Delta region, China. *Environ. Pollut.* **2017**, *226*, 143–153. [CrossRef]
14. Stafoggia, M.; Schwartz, J.; Badaloni, C.; Bellander, T.; Alessandrini, E.; Cattani, G.; de' Donato, F.; Gaeta, A.; Leone, G.; Lyapustin, A.; et al. Estimation of daily PM10 concentrations in Italy (2006–2012) using finely resolved satellite data, land use variables and meteorology. *Environ. Int.* **2017**, *99*, 234–244. [CrossRef]
15. Shi, Y.; Lau, K.K.-L.; Ng, E. Incorporating wind availability into land use regression modelling of air quality in mountainous high-density urban environment. *Environ. Res.* **2017**, *157*, 17–29. [CrossRef]
16. Son, Y.; Osornio-Vargas, Á.R.; O'Neill, M.S.; Hystad, P.; Texcalac-Sangrador, J.L.; Ohman-Strickland, P.; Meng, Q.; Schwander, S. Land use regression models to assess air pollution exposure in Mexico City using finer spatial and temporal input parameters. *Sci. Total Environ.* **2018**, *639*, 40–48. [CrossRef]
17. Zou, B.; Chen, J.; Zhai, L.; Fang, X.; Zheng, Z.; Zou, B.; Chen, J.; Zhai, L.; Fang, X.; Zheng, Z. Satellite Based Mapping of Ground PM2.5 Concentration Using Generalized Additive Modeling. *Remote Sens.* **2016**, *9*, 1. [CrossRef]
18. Wu, C.-D.; Chen, Y.-C.; Pan, W.-C.; Zeng, Y.-T.; Chen, M.-J.; Guo, Y.L.; Lung, S.-C.C. Land-use regression with long-term satellite-based greenness index and culture-specific sources to model PM2.5 spatial-temporal variability. *Environ. Pollut.* **2017**, *224*, 148–157. [CrossRef]
19. He, J.; Zha, Y.; Zhang, J.; Gao, J. Aerosol indices derived from MODIS data for indicating aerosol-induced air pollution. *Remote Sens.* **2014**, *6*, 1587–1604. [CrossRef]

20. Just, A.; De Carli, M.; Shtein, A.; Dorman, M.; Lyapustin, A.; Kloog, I.; Just, A.C.; De Carli, M.M.; Shtein, A.; Dorman, M.; et al. Correcting Measurement Error in Satellite Aerosol Optical Depth with Machine Learning for Modeling PM2.5 in the Northeastern USA. *Remote Sens.* **2018**, *10*, 803. [CrossRef]
21. Wan, Z. *MODIS Land Surface Temperature Products Users' Guide*; Institute for Computational Earth System Science, University of California: Santa Barbara, CA, USA, 2006.
22. U.S. Geological Survey. *Landsat—Earth Observation Satellites*; Version 1.; U.S. Geological Survey: Reston, VA, USA, 2015; Volume 2015–3081.
23. Olmanson, L.G.; Brezonik, P.L.; Finlay, J.C.; Bauer, M.E. Comparison of Landsat 8 and Landsat 7 for regional measurements of CDOM and water clarity in lakes. *Remote Sens. Environ.* **2016**, *185*, 119–128. [CrossRef]
24. Bilal, M.; Nichol, J.E.; Bleiweiss, M.P.; Dubois, D. A Simplified high resolution MODIS aerosol retrieval algorithm (SARA) for use over mixed surfaces. *Remote Sens. Environ.* **2013**, *136*, 135–145. [CrossRef]
25. Meng, X.; Fu, Q.; Ma, Z.; Chen, L.; Zou, B.; Zhang, Y.; Xue, W.; Wang, J.; Wang, D.; Kan, H.; et al. Estimating ground-level PM10 in a Chinese city by combining satellite data, meteorological information and a land use regression model. *Environ. Pollut.* **2016**, *208*, 177–184. [CrossRef]
26. Shahraiyni, H.T.; Sodoudi, S. Statistical modeling approaches for pm10prediction in urban areas; A review of 21st-century studies. *Atmosphere* **2016**, *7*, 15. [CrossRef]
27. Vermote, E.; Justice, C.; Claverie, M.; Franch, B. Preliminary analysis of the performance of the Landsat 8/OLI land surface reflectance product. *Remote Sens. Environ.* **2016**, *185*, 46–56. [CrossRef]
28. Ayres-Sampaio, D.; Teodoro, A.C.; Sillero, N.; Santos, C.; Fonseca, J.; Freitas, A. An investigation of the environmental determinants of asthma hospitalizations: An applied spatial approach. *Appl. Geogr.* **2014**, *47*, 10–19. [CrossRef]
29. Naughton, O.; Donnelly, A.; Nolan, P.; Pilla, F.; Misstear, B.D.; Broderick, B. A land use regression model for explaining spatial variation in air pollution levels using a wind sector based approach. *Sci. Total Environ.* **2018**, *630*, 1324–1334. [CrossRef]
30. Li, X.; Zhang, Y.; Bao, Y.; Luo, J.; Jin, X.; Xu, X.; Song, X.; Yang, G. Exploring the Best Hyperspectral Features for LAI Estimation Using Partial Least Squares Regression. *Remote Sens.* **2014**, *6*, 6221–6241. [CrossRef]
31. Chen, G.; Meentemeyer, R. Remote Sensing of Forest Damage by Diseases and Insects. In *Remote Sensing for Sustainability*; Weng, Q., Ed.; Remote Sensing Applications Series; CRC Press: Boca Raton, FL, USA, 2016; p. 357. ISBN 9781315354644.
32. Xu, W.; Riley, E.A.; Austin, E.; Sasakura, M.; Schaal, L.; Gould, T.R.; Hartin, K.; Simpson, C.D.; Sampson, P.D.; Yost, M.G.; et al. Use of mobile and passive badge air monitoring data for NO X and ozone air pollution spatial exposure prediction models. *J. Expo. Sci. Environ. Epidemiol.* **2017**, *27*, 184–192. [CrossRef]
33. Rosero-Vlasova, O.A.; Vlassova, L.; Pérez-Cabello, F.; Montorio, R.; Nadal-Romero, E. Modeling soil organic matter and texture from satellite data in areas affected by wildfires and cropland abandonment in Aragón, Northern Spain. *J. Appl. Remote Sens.* **2018**, *12*, 1. [CrossRef]
34. Alvarez-Mendoza, C.I.; Teodoro, A.; Ramirez-Cando, L. Spatial estimation of surface ozone concentrations in Quito Ecuador with remote sensing data, air pollution measurements and meteorological variables. *Environ. Monit. Assess.* **2019**, *191*, 155. [CrossRef]
35. Liu, W.; Li, X.; Chen, Z.; Zeng, G.; León, T.; Liang, J.; Huang, G.; Gao, Z.; Jiao, S.; He, X.; et al. Land use regression models coupled with meteorology to model spatial and temporal variability of NO2 and PM10 in Changsha, China. *Atmos. Environ.* **2015**, *116*, 272–280. [CrossRef]
36. Gardner, M.; Dorling, S. Artificial neural networks (the multilayer perceptron)—A review of applications in the atmospheric sciences. *Atmos. Environ.* **1998**, *32*, 2627–2636. [CrossRef]
37. Secretaria del Ambiente de Quito Red Metropolitana de Monitoreo Atmosférico de Quito. Available online: http://www.quitoambiente.gob.ec/ambiente/index.php/generalidades (accessed on 26 June 2018).
38. Alvarez, C.I.; Teodoro, A.; Tierra, A. Evaluation of automatic cloud removal method for high elevation areas in Landsat 8 OLI images to improve environmental indexes computation. In Proceedings of the SPIE 10428, Earth Resources and Environmental Remote Sensing/GIS Applications VIII 1042809, Warsaw, Poland, 5 October 2017; Volume 10428, pp. 1042809–1042812.
39. Alvarez-Mendoza, C.I.; Teodoro, A.; Ramirez-Cando, L. Improving NDVI by removing cirrus clouds with optical remote sensing data from Landsat-8—A case study in Quito, Ecuador. *Remote Sens. Appl. Soc. Environ.* **2019**, *13*, 257–274. [CrossRef]

40. Othman, N.; Jafri, M.Z.M.; San, L.H. Estimating particulate matter concentration over arid region using satellite remote sensing: A case study in Makkah, Saudi Arabia. *Mod. Appl. Sci.* **2010**, *4*, 131. [CrossRef]
41. Bilguunmaa, M.; Batbayar, J.; Tuya, S. Estimation of PM10 concentration using satellite data in Ulaanbaatar City. *SPIE Asia Pac. Remote Sens.* **2014**, 92591O. [CrossRef]
42. Ángel, M.; Gutiérrez, R. Uso de Modelos Lineales Generalizados (MLG) para la interpolación espacial de PM10 utilizando imágenes satelitales Landsat para la ciudad de Bogotá, Colombia. *Perspectiva Geográfica.* **2017**, *22*, 105–121. [CrossRef]
43. Lee, J.H.; Ryu, J.E.; Chung, H.I.; Choi, Y.Y.; Jeon, S.W.; Kim, S.H. Development of spatial scaling technique of forest health sample point information. In Proceedings of the International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences—ISPRS Archives, Beijing, China, 30 April 2018; Volume 42, pp. 751–756.
44. Ghaleb, F.; Mario, M.; Sandra, A. Regional Landsat-Based Drought Monitoring from 1982 to 2014. *Climate* **2015**, *3*, 563–577. [CrossRef]
45. Sobrino, J.A.; Jiménez-Muñoz, J.C.; Sòria, G.; Romaguera, M.; Guanter, L.; Moreno, J.; Plaza, A.; Martínez, P. Land surface emissivity retrieval from different VNIR and TIR sensors. *IEEE Trans. Geosci. Remote Sens.* **2008**, *46*, 316–327. [CrossRef]
46. Li, S.; Jiang, G.M. Land Surface Temperature Retrieval from Landsat-8 Data with the Generalized Split-Window Algorithm. *IEEE Access* **2018**, *6*, 18149–18162. [CrossRef]
47. Habermann, M.; Billger, M.; Haeger-Eugensson, M. Land use Regression as Method to Model Air Pollution. Previous Results for Gothenburg/Sweden. *Procedia Eng.* **2015**, *115*, 21–28. [CrossRef]
48. Zhang, J.J.Y.; Sun, L.; Barrett, O.; Bertazzon, S.; Underwood, F.E.; Johnson, M. Development of land-use regression models for metals associated with airborne particulate matter in a North American city. *Atmos. Environ.* **2015**, *106*, 165–177. [CrossRef]
49. Williams, L.J.; Abdi, H.; Williams, L.J. Partial Least Squares Methods: Partial Least Squares Correlation and Partial Least Square Regression. In *Computational Toxicology: Volume II.*; Reisfeld, B., Mayeno, A.N., Eds.; Humana Press: Totowa, NJ, USA, 2013; Volume 930, pp. 549–579. ISBN 978-1-62703-059-5.
50. Sheela, K.G.; Deepa, S.N. Review on Methods to Fix Number of Hidden Neurons in Neural Networks. *Math. Probl. Eng.* **2013**, *2013*, 1–11. [CrossRef]
51. Cattani, G.; Gaeta, A.; Di Menno di Bucchianico, A.; De Santis, A.; Gaddi, R.; Cusano, M.; Ancona, C.; Badaloni, C.; Forastiere, F.; Gariazzo, C.; et al. Development of land-use regression models for exposure assessment to ultrafine particles in Rome, Italy. *Atmos. Environ.* **2017**, *156*, 52–60. [CrossRef]
52. Wang, M.; Sampson, P.D.; Hu, J.; Kleeman, M.; Keller, J.P.; Olives, C.; Szpiro, A.A.; Vedal, S.; Kaufman, J.D. Combining Land-Use Regression and Chemical Transport Modeling in a Spatiotemporal Geostatistical Model for Ozone and PM 2.5. *Environ. Sci. Technol.* **2016**, *50*, 5111–5118. [CrossRef]
53. Alexeeff, S.E.; Schwartz, J.; Kloog, I.; Chudnovsky, A.; Koutrakis, P.; Coull, B.A. Consequences of kriging and land use regression for PM2.5 predictions in epidemiologic analyses: Insights into spatial variability using high-resolution satellite data. *J. Expo. Sci. Environ. Epidemiol.* **2015**, *25*, 138–144. [CrossRef]
54. Beloconi, A.; Chrysoulakis, N.; Lyapustin, A.; Utzinger, J.; Vounatsou, P. Bayesian geostatistical modelling of PM10 and PM2.5 surface level concentrations in Europe using high-resolution satellite-derived products. *Environ. Int.* **2018**, *121*, 57–70. [CrossRef]
55. Remer, L.A.; Mattoo, S.; Levy, R.C.; Munchak, L.A. MODIS 3 km aerosol product: Algorithm and global perspective. *Atmos. Meas. Tech.* **2013**, *6*, 1829–1844. [CrossRef]
56. Teodoro, A. A study on the Quality of the Vegetation Index obtainded from MODIS Data. In Proceedings of the 2015 IEEE International Geoscience and Remote Sensing Symposium, Milan, Italy, 26–31 July 2015; pp. 3365–3368.
57. Saucy, A.; Röösli, M.; Künzli, N.; Tsai, M.Y.; Sieber, C.; Olaniyan, T.; Baatjies, R.; Jeebhay, M.; Davey, M.; Flückiger, B.; et al. Land use regression modelling of outdoor NO2 and PM2.5 concentrations in three low income areas in the western cape province, South Africa. *Int. J. Environ. Res. Public Health* **2018**, *15*, 1452. [CrossRef]
58. Lv, Y.; Liu, J.; Yang, T. Nonlinear PLS Integrated with Error-Based LSSVM and Its Application to NO_2 Modeling. *Ind. Eng. Chem. Res.* **2012**, *51*, 16092–16100. [CrossRef]
59. Secretaria del Ambiente de Quito. *IAMQ/18*; Secretaria del Ambiente de Quito: Quito, Ecuador, 2018.

60. Romero, D.; El parque automotor aumenta y complica más la movilidad. El Comer. *2017, 1.* Available online: https://www.elcomercio.com/actualidad/aumento-parque-automotor-quito-movilidad.html (accessed on 13 June 2019).
61. Todoroski Air Sciences. *Air Quality Impact Assessment Sandy Point Quarry Epl Variation*; Todoroski Air Sciences: Eastwood, Australia, 2019.

Environ Monit Assess (2019) 191: 155
https://doi.org/10.1007/s10661-019-7286-6

Spatial estimation of surface ozone concentrations in Quito Ecuador with remote sensing data, air pollution measurements and meteorological variables

Cesar I. Alvarez-Mendoza · Ana Teodoro · Lenin Ramirez-Cando

Received: 5 September 2018 / Accepted: 30 January 2019 / Published online: 11 February 2019

Abstract Surface ozone is problematic to air pollution. It influences respiratory health. The air quality monitoring stations measure pollutants as surface ozone, but they are sometimes insufficient or do not have an adequate distribution for understanding the spatial distribution of pollutants in an urban area. In recent years, some projects have found a connection between remote sensing, air quality and health data. In this study, we apply an empirical land use regression (LUR) model to retrieve surface ozone in Quito. The model considers remote sensing data, air pollution measurements and meteorological variables. The objective is to use all available Landsat 8 images from 2014 and the air quality monitoring station data during the same dates of image acquisition. Nineteen input variables were considered, selecting by a stepwise regression and modelling with a partial least square (PLS) regression to avoid multicollinearity. The final surface ozone model includes ten independent variables and presents a coefficient of determination (R^2) of 0.768. The model proposed help to understand the spatial concentration of surface ozone in Quito with a better spatial resolution.

Keywords Landsat 8 · Quito · Ozone · PLS · Air modelling

C. I. Alvarez-Mendoza · A. Teodoro
Department of Geosciences, Environment and Land Planning, Faculty of Sciences, University of Porto, Rua Campo Alegre 687, 4169-007 Porto, Portugal

A. Teodoro
e-mail: amteodor@fc.up.pt

C. I. Alvarez-Mendoza (✉) · L. Ramirez-Cando
Grupo de Investigación Ambiental en el Desarrollo Sustentable GIADES, Carrera de Ingeniería Ambiental, Universidad Politécnica Salesiana, Quito, Ecuador
e-mail: calvarezm@ups.edu.ec

L. Ramirez-Cando
e-mail: lramirez@ups.edu.ec

A. Teodoro
Earth Sciences Institute (ICT), Pole of the FCUP, University of Porto, Porto, Portugal

Introduction

Surface ozone (O_3) is one of the principal greenhouse gases (US Department of Commerce 2018). It is produced in the troposphere and is not emitted directly into the air. A chemical reaction between nitrogen oxides (NO_x), volatile organic compounds (VOC) and sunlight produce O_3 (US EPA 2014). Thus, urban growth, vehicular traffic and industry are sources of NO_x and VOC in cities, deteriorating the vegetation conditions (Monks et al. 2015), the air quality and creating a health problem (US EPA; WHO (World 2013)).

Several cities around the world have an air quality monitoring network (AQMN) to manage air pollution (Liang et al. 2016; Lee et al. 2018). One of the cities with an AQMN is Quito, the capital of Ecuador. The city has traffic and population problems that increase air pollution. Its AQMN is the "Red Metropolitana de

Monitoreo Atmosférico de Quito" (REMMAQ), constituted by nine stations. It has managed the air quality in Quito in real time since 2002 (Secretaria del Ambiente de Quito 2018). The REMMAQ stations measure air pollutants such as carbon monoxide (CO), nitrogen dioxide (NO_2) as part of NO_x, sulphur dioxide (SO_2), particulate matter less than 10 μm (PM10), fine particles less than 2.5 μm (PM2.5) and O_3. Nevertheless, the number of stations is insufficient to measure the air quality in all urban zones in the city.

Some empirical models to retrieve the spatial concentration of air pollutants have been developed using variables such as road information and vegetation. The land use regression (LUR) models are the basis of most of these approaches. The principle of LUR focuses on the environmental characteristics of the place where the pollutant is present (Habermann et al. 2015). Some models consider remote sensing data, meteorological data (MD), aerosol optical depth (AOD) field measurements and AQMN data (Liu et al. 2007; Chen et al. 2010; Zhang et al. 2018). In most of these studies, the limitations are related to the input variables, especially AOD field measurements. This is because models require AOD parameters to obtain high-resolution spatialization (Bilal et al. 2013; Zhang et al. 2018). The most commonly used remote sensing data are Landsat (Chen et al. 2014; Meng et al. 2015; Zheng et al. 2017) and MODIS (Stafoggia et al. 2017; Braun et al. 2018) sensors. The main advantage of Landsat images in specific Landsat 8 (U.S. Geological Survey 2016) is the high spatial resolution to map middle cities. Their limitation is the temporal resolution (16 days) (U.S. Geological Survey 2016). The advantage of MODIS is its high temporal resolution, but the major limitation is the low spatial resolution, which limits the accurate retrieval of maps (Daac et al. 2012). Moreover, remote sensing data are used to obtain environmental variables such as vegetation health (Jia et al. 2014; Zhang et al. 2016) to input variables in the air pollutant models. Furthermore, empirical models using remote sensing data are focused on only some air pollutants, such as NO_2, PM10 and PM2.5. At present, the main challenge is to retrieve the remaining air pollutants, such as O_3, which is considered only in few studies (e.g. Mok et al. 2018).

In the case of Quito, a study found the spatial distribution of PM10 by applying remote sensing data (Alvarez and Padilla Almeida 2016). The main limitation of the study was the small quantity of data used (three images). On the other hand, a study making a comparison between remote sensing to retrieve air pollutant in Quito is considered (Alvarez-Mendoza et al. 2018b). However, there are few studies about air quality in the city, specifically considering O_3 (Cazorla 2016). Thus, the possibility of obtaining AQMN public data, and combining them with other environmental variables, can lead to new models for retrieving air pollutants in places where AQMN are insufficient.

This study uses remote sensing data, air pollution measurements and meteorological variables to retrieve O_3 for 1 year (2014) in Quito. Moreover, this study combines two regression techniques, stepwise regression (SWR) and partial least-square (PLS) regression, to compute the O_3 model, finding the fittest model to spatialize the variable in all the areas. The main objective is to find the spatial variables that influence O_3 in Quito.

Materials and methods

Study area

This study was developed in Quito, the capital of Ecuador. The city elevation is approximately 2800 m above sea level. During 2014, the mean minimum and maximum temperatures were 9.0 °C and 25.4 °C (Instituto Nacional de Meteorología e Hidrología 2016). Furthermore, Quito has a dry season and a wet season. It does not have four seasons considering that the city is in the middle of the tropic zone. The latitude and longitude of the study area are 0° 30′ S to 0° 10′ N and 78° 10′ W to 78° 40′ W. These coordinates delimit most of the urban zone, which is divided into urban parishes (Fig. 1).

Air pollutant ground data

The daily air pollutant concentration data from 2014 were obtained from the REMMAQ stations. The REMMAQ has nine automatic stations that have been operated by the "Secretaria del Ambiente de Quito" since 2002 (Fig. 1). The stations measure concentrations of air pollutants such as PM2.5, SO_2, CO, O_3, NO_2, PM10 and MD (Table 1). In this study, daily average measurements were considered to match with the satellite overpass (Fig. 2) (See section 2.4). Furthermore, only complete datasets were used, which means that if a dataset was incomplete, it was not considered for the

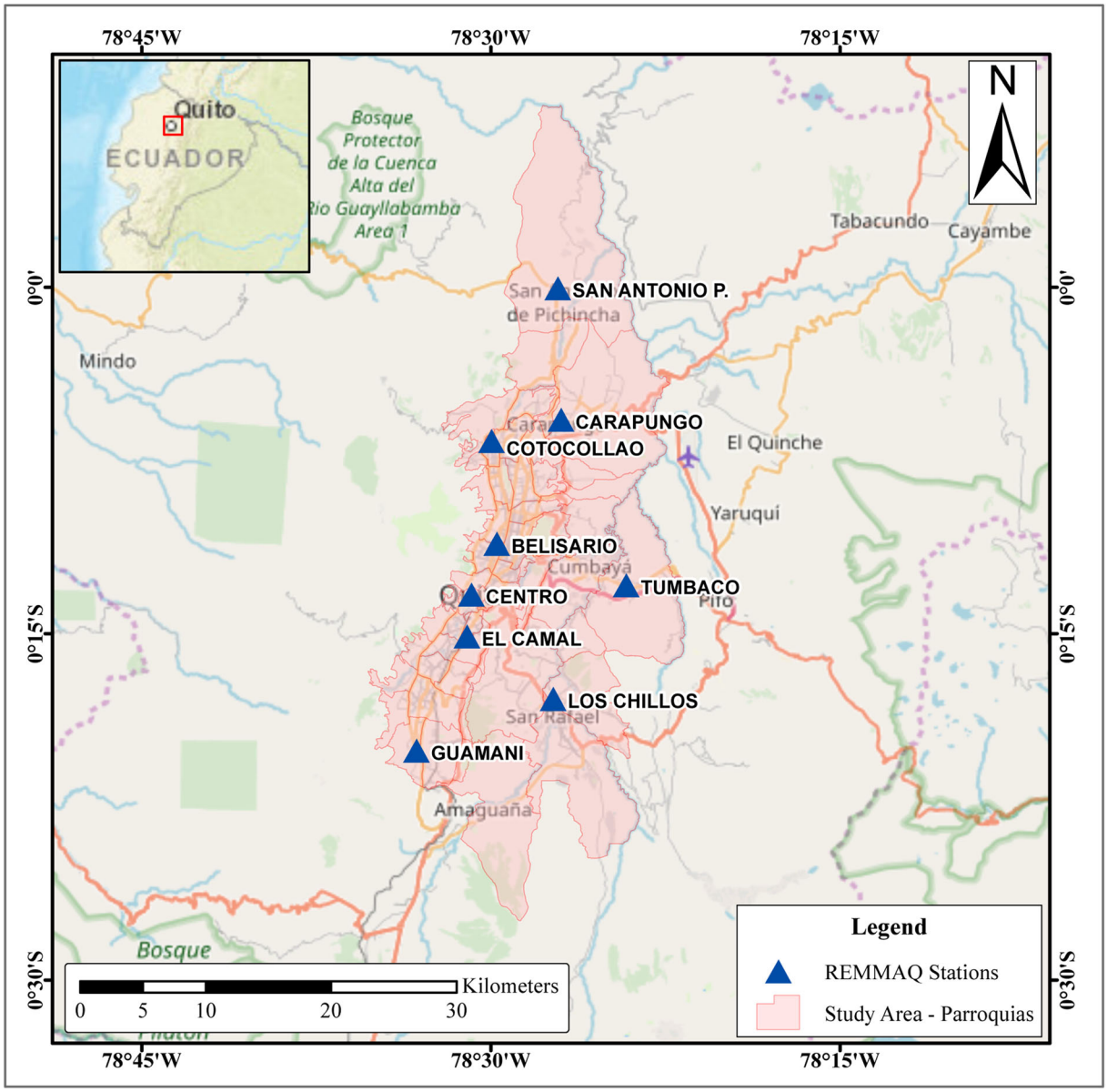


Fig. 1 Quito's urban parishes were considered as the study area. The blue marks represent the REMMAQ stations

model establishment. PM2.5, SO_2, CO, and NO_2 were the complete datasets to estimate O_3. The pollutant concentration was measured in micrograms per cubic metre ($\mu g/m^3$) according to the Environmental Protection Agency (EPA) methods. The O_3 measuring device was a Teledyne API/T400, and the collection method was EPA No. EQOA-0992-087 (Secretaria del Ambiente de Quito 2018). The hourly pollutant concentration data have public access (http://www.quitoambiente.gob.ec/ambiente/index.php/datos-horarios-historicos).

Meteorological data

The MD were collected only by eight REMMAQ stations (Table 1). The data used were the daily average temperature (TMP) in degrees Celsius (°C), relative humidity (HM) in percentage (%) and solar radiation (SR) in watt per square metre (W/m^2). The precipitation measurements were not used because most of the values were null in the time range considered.

In both cases (air pollutant ground data and meteorological data), the R software was used to analyse the

Table 1 Field sensors of the REEMAQ

Station	Variables measured
Cotocollao	PM2.5, SO_2, CO, O_3, NO_2, PM10, MD
Carcelen	PM2.5, SO_2, CO, O_3, NO_2, PM10, MD
Belisario	PM2.5, SO_2, CO, O_3, NO_2, MD
Jipijapa	PM2.5, SO_2, CO, O_3, NO_2, PM10, MD
Camal	PM2.5, SO_2, CO, O_3, NO_2, MD
Centro	PM2.5, SO_2, CO, O_3, NO_2
Guamani	SO_2, CO, O_3, NO_2, PM10, MD
Tumbaco	SO_2, O_3, PM10, MD
Los Chillos	PM2.5, SO_2, CO, O_3, NO_2, MD

data and compute the statistics. The packages *readxl* and *stringi* were used.

Remote sensing data

Landsat 8 is a satellite launched on February 11, 2013. It is the last satellite of the Landsat project launched. The satellite carries two push-brown instruments to collect land remote sensing data on an image: the Operational Land Imager (OLI) with nine bands and the thermal infrared sensors (TIRS) with two bands. Additionally, the Landsat 8 data file provides a quality assessment (QA) band to assess the different image products. The Landsat 8 images are freely available on the United States Geological Survey (USGS) website. The USGS develops research-quality and application-ready products such as the Landsat 8 Surface Reflectance Level-2 products (L2T). These products are generated from the Landsat Surface Reflectance Code (LaSRC) (Vermote et al. 2016). The LaSRC products are radiometric and atmospherically corrected. The LaSRC products include surface reflectance of the OLI bands (bands 1 to 9), top-of-atmosphere brightness temperature (BT) (band 10 and band 11) and some environmental indexes such as the normalised difference vegetation index (NDVI), soil-adjusted vegetation index (SAVI) and enhanced vegetation index (EVI).

In this study, Landsat 8 L2T images were downloaded from the Earth Resources Observation and Science (EROS) Center Science Processing Architecture (ESPA) at the demand interface (https://espa.cr.usgs.gov/). The search criteria were images in 2014 with less than 20% cloud cover in the study area. One of the challenges was to choose the subset of images without high cloud density in the study area (Alvarez-Mendoza et al. 2018a). According to the search criteria, ten images (path 11; row 60) were selected (Table 2).

Considering the direct influence of the sunlight over O_3 concentration (US EPA 2014) and knowing the principle of passive remote sensing data to capture the radiation-measured reflectance sunlight (Liew 2001; NASA EOSDIS 2018), bands 1 to 7 (visible and infrared bands) (U.S. Geological Survey 2016) were used as input variables. NDVI, SAVI and EVI were used to highlight the vegetation because there is a high relation

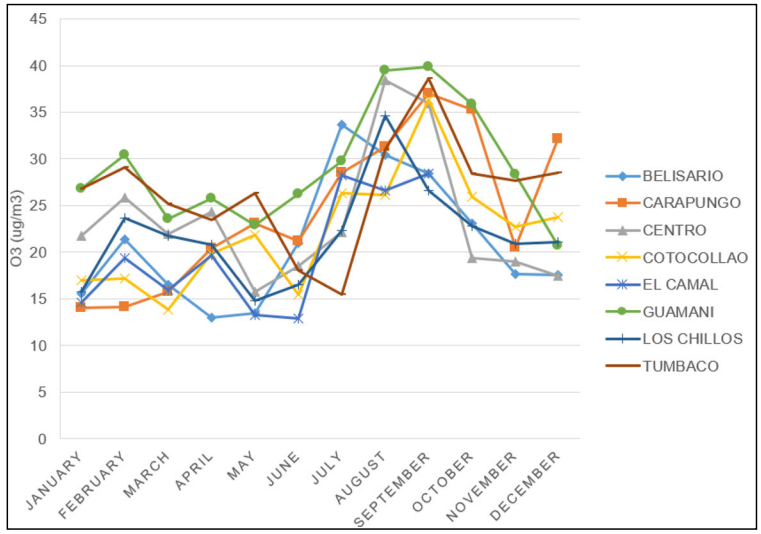


Fig. 2 Mean levels from 10:00 to 11:00 (GMT-5) of O_3 concentration ($\mu g/m^3$) observed in each month during 2014. The San Antonio P. station did not present measures during 2014

Table 2 Landsat 8 L2T images selected

No.	Image	Date
1	LC08_L1TP_010060_20140115_20170426_01_T1	15/01/2014
2	LC08_L1TP_010060_20140131_20170426_01_T1	31/01/2014
3	LC08_L1TP_010060_20140216_20170425_01_T1	16/02/2014
4	LC08_L1TP_010060_20140304_20170425_01_T1	04/03/2014
5	LC08_L1TP_010060_20140405_20170424_01_T1	05/04/2014
6	LC08_L1TP_010060_20140608_20170422_01_T1	08/06/2014
7	LC08_L1TP_010060_20140710_20170421_01_T1	10/07/2014
8	LC08_L1TP_010060_20140726_20170420_01_T1	26/07/2014
9	LC08_L1TP_010060_20140811_20170420_01_T1	11/08/2014
10	LC08_L1TP_010060_20141030_20170418_01_T1	30/10/2014

between O_3 and vegetation (Sicard et al. 2017). The indexes were obtained from LaSRC and multiplied by 0.0001 (USGS 2017) to retrieve the surface environmental indexes (values between − 1 and 1).

The NVDI provides information about health vegetation, using band 4 (B4) and band 5 (B5) in Landsat 8 images. It is computed using Eq. 1.

$$\mathrm{NDVI} = \frac{\mathrm{B5}-\mathrm{B4}}{\mathrm{B5}+\mathrm{B4}} \tag{1}$$

The SAVI is an improvement of NDVI considering a soil correction factor (usually LS = 0.5). Considering Landsat 8, it uses B4 and B5 as input (Eq. 2).

$$\mathrm{SAVI} = (1+\mathrm{LS})\,\frac{\mathrm{B5}-\mathrm{B4}}{\mathrm{B5}+\mathrm{B4}+\mathrm{LS}} \tag{2}$$

The EVI enhances the vegetation in areas with high biomass. Thus, EVI helps to identify stress vegetation using Eq. 3.

$$\mathrm{EVI} = G \times \frac{\mathrm{B5}-\mathrm{B4}}{\mathrm{B5}+\mathrm{C1}\times\mathrm{B4}-\mathrm{C2}\times\mathrm{B2}+L} \tag{3}$$

Where the gain factor (G) is 2.5, L is the canopy background adjustment ($L = 1$) and C1 and C2 are coefficients for atmospheric resistance (C1 = 6, C2 = 7.5). The B4 and B5 have a high contrast in the detection of built-up areas and bare lands areas (Assyakur et al. 2012).

Moreover, the land surface temperature (LST) retrieved from remote sensing has been used in other studies to estimate the air quality (Chen et al. 2014). It was computed as a function of BT. Equation 4 represents the LST in degrees Celsius.

$$\mathrm{LST} = \frac{\mathrm{BT}}{\left(1+\left(\frac{\lambda\times\mathrm{BT}}{p}\right)\ln E\right)} - 273.15 \tag{4}$$

Where λ is the centre wavelength ($\lambda = 10.8$ μm), ρ is a constant obtained in Eq. 5, E is the emissivity as Eq. 6 and 273.15 is the value to transform degrees Kelvin to degrees Celsius.

The constant ρ is estimated using Eq. 5, where h is the Planck constant (6.626e-34 Js), c is the speed of light (2.998e8 m/s), and s is the Boltzmann constant (1.38e-23 J/K).

$$p = \frac{h\times c}{s} \tag{5}$$

Equation 6 represents the emissivity E (Vieira et al. 2016). E is the efficiency of a surface that emits heat as thermal infrared (TIR) radiation (Gillespie 2014).

$$E = \begin{cases} E_s, \mathrm{NDVI} < \mathrm{NDVI_s} \\ E_s + (E_v - E_s)\,P_V, \quad \mathrm{NDVI_s} \le \mathrm{NDVI} \le \mathrm{NDVI_v} \\ E_v, \quad \mathrm{NDVI} > \mathrm{NDVI_v} \end{cases} \tag{6}$$

Where E_s represents the emissivity for soil. A value of 0.973 is used in this study (Sobrino et al. 2008). E_v is the vegetation emissivity with a value of 0.985 in this study (Sobrino et al. 2008). $\mathrm{NDVI_v}$ is the NDVI in vegetation with a value of 0.2 (Vieira et al. 2016), $\mathrm{NDVI_s}$ is the NDVI in the soil with a value of 0.5 (Vieira et al. 2016) and Pv is the proportion of vegetation in the area using Eq. 7.

$$P_V = \left(\frac{NDVI - NDVI_s}{NDVI_v - NDVI_s}\right)^2 \quad (7)$$

The remote sensing variables were represented as raster data (GeoTIFF format). They were computed in R studio software with the *rgdal* and *raster* packages. Through the shapefile of REMMAQ stations, the raster values for each station were extracted. The package *dismo* was used to perform this task.

Model building

The first step in building the model is the compilation of all possible variables (air measurement data, meteorological data and remote sensing data) in a database. Each row in the table has all the values of these variables in a REMMAQ station during the date established (Table 3).

LUR models are a good alternative for finding the spatial location of pollutants (Larkin et al. 2017). LUR models are empirical regression models that consider the pollutant of interest as the dependent variable and other geographical variables as independent variables (meteorological data, traffic, topography, remote sensing data, etc.). In this study, we generate an LUR model using the available data from each station on different dates during 2014 to preserve the accuracy of the variables.

Assuming that multicollinearity between variables is real, especially between remote sensing variables (Chen and Meentemeyer 2016), a preliminary correlation analysis was realised to provide an overview of which variables are more adequate for integration into the model.

To select the fittest predictor variables and the best model to predict O_3, a subset analysis is performed with stepwise regression. The subset analysis used four analyses: the residual sum of squares for each model (RSS), the adjusted regression coefficient R^2 (Adj. R^2), Mallows' Cp (CP) and the Bayesian information criterion (BIC). The R-package used to compute this was *leaps*.

The original LUR model with all the possible predictor variables as input in the analysis is shown in Eq. 8.

$$\begin{aligned} O_3 = {} & a\mathrm{PM2.5} + b\mathrm{SO_2} + c\mathrm{CO} + d\mathrm{NO_2} + e\mathrm{TMP} \\ & + f\mathrm{HUM} + g\mathrm{SR} + h\mathrm{B1} + i\mathrm{B2} + j\mathrm{B3} \\ & + k\mathrm{B4} + l\mathrm{B5} + m\mathrm{B6} + n\mathrm{B7} + o\mathrm{NDVI} \\ & + p\mathrm{SAVI} + q\mathrm{EVI} + r\mathrm{LST} + I \end{aligned} \quad (8)$$

Where $a, b, c \ldots r$ are the coefficients of the regression model, and I is the intercept in the equation. The subset analysis reduces the number of input variables with the considered criteria (RSS, Adj. R^2, CP, BIC).

Once the input variables are selected, a PLS regression is applied to avoid the multicollinearity between the variable subsets. PLS is a technique applied in cases where traditional regression models fail, and the predictors have a high correlation, as shown in Eqs. 9 and 10.

$$X = TP^T + E \quad (9)$$

$$Y = UQ^T + \mathrm{F} \quad (10)$$

Where X is a $n\ x\ m$ matrix of predictors, Y is a $n\ x\ p$ matrix of responses; T and U are $n\ x\ l$ matrices that are, respectively, projections of X and projections of Y; P and Q are, respectively $m\ x\ l$ and $p\ x\ l$ orthogonal loading matrices; and matrices E and F are the error terms. The decompositions of X and Y are made in order to

Table 3 Variables considered in the model

No.	Variable	Units
Air pollutants ground data	O_3, PM2.5, SO_2, CO, NO_2	$\mu g/m^3$
Meteorological data	Temperature (TMP)	°C
	Relative humidity (HUM)	%
	Solar radiation (SR)	W/m^2
Remote sensing data	Band 1 (B1), Band 2 (B2), Band 3 (B3), Band 4 (B4), Band 5 (B5), Band 6 (B6), Band 7 (B7)	Surface reflectance
	Environmental indexes: NDVI, SAVI, EVI	–
	Land surface temperature (LST)	°C

maximise the covariance between T and U. Additionally, PLS generates an orthogonal transformation to obtain components by finding the most appropriate model to explain the variance starting from the maximised covariance matrixes (Williams et al. 2013). In the case of remote sensing data, some studies consider multicollinearity when the same sensor is used to obtain different variables (Chen and Meentemeyer 2016; Gholizadeh and Robeson 2016). Finally, the validation is performed by cross-validation (Fig. 3) and the criterion to accept or reject models where R^2, RMSE, predicated vs measured graphic and residuals analysis. The R-packages used were *pls* and *plsdepot*.

Results and discussion

Building the ozone LUR model

The LUR model tested 19 variables (18 independent variables or predictors and O_3 as the dependent variable), matching all variables (air measurement data, meteorological data and remote sensing data). The result is a database with 36 observations, where most of the remote sensing data variables show a high correlation (Fig. 4). The high correlation or multicollinearity (in some cases near 1) indicates that some variables are highly related, such as NDVI, SAVI and EVI, or the visible bands (B1, B2, B3, B4). On the other hand, the highest correlation between all predictors with O_3 is PM2.5, showing a value of − 0.44. The highest correlation considering only the remote sensing data variables is B6 with 0.22.

To find the model with the best fit, a stepwise regression subset is used. In the first instance (Fig. 5), the coefficient of determination (R^2) is near 0.68, considering all 18 independent variables to build the model. The subset variables are analysed by the less Akaike information criterion (AIC) and the maximum Adj. R^2.

The preliminary predictors are known (Fig. 5); so to find a simple model with fewer input variables, a new subset of variables, applying RSS, Adj. R^2, CP and BIC

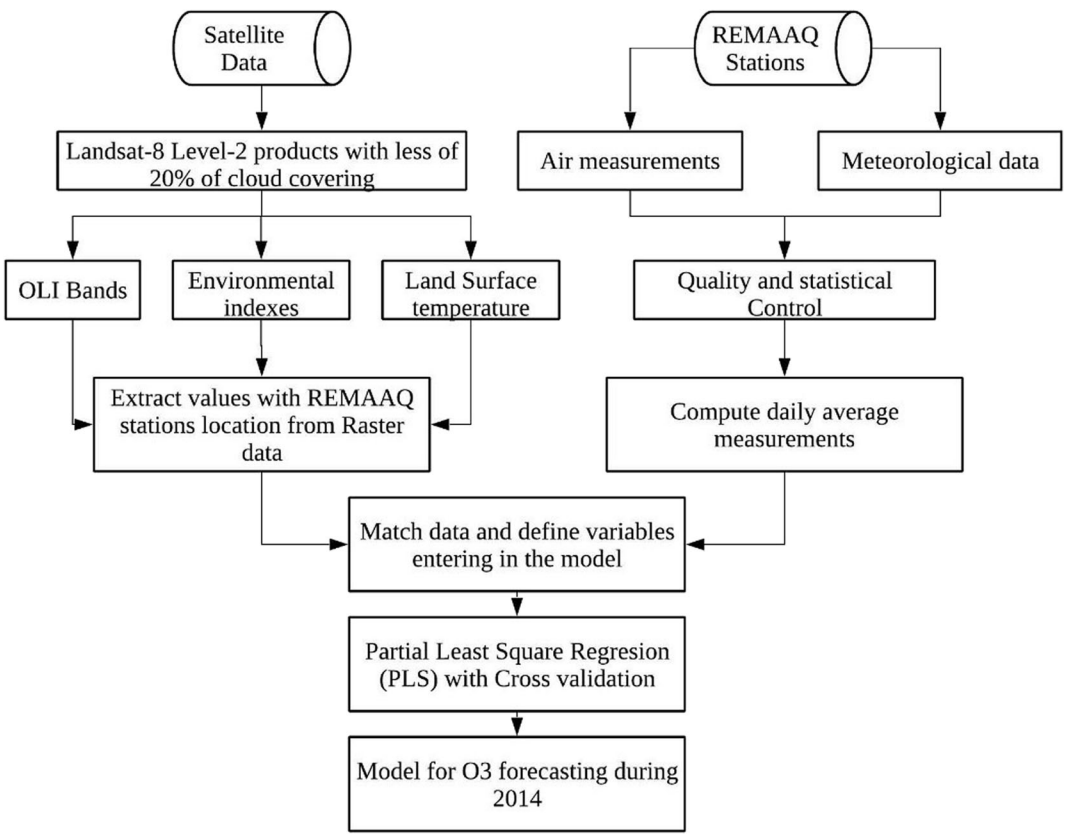


Fig. 3 Methodology workflow

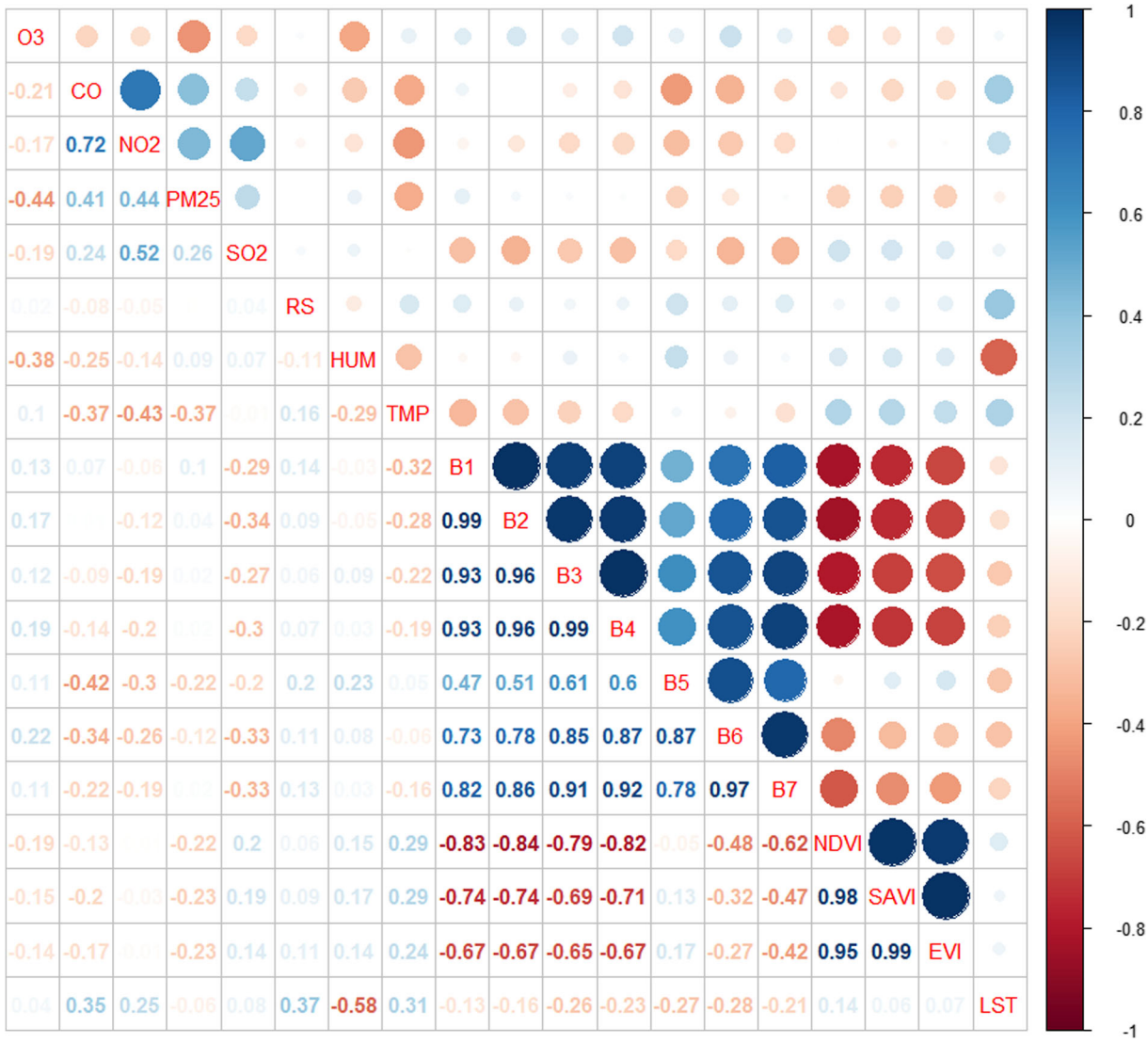


Fig. 4 Correlation graph between input variables

criteria, are analysed (Fig. 6). Analysing the four criteria, 11 independent variables are used to build the simplest model (PM2.5, HUM, TMP, B2, B4, B5, B7, NDVI, SAVI, EVI).

The 11 variables chosen were then considered in the PLS analysis (Fig. 7). The number of components in PLS regression was nine. These components explain most of the percentage of variance (Table 4), after cross-validation (data not shown). The R^2 obtained was 0.77, and the RMSE was 3.03 through the PLS regression.

Avoiding the multicollinearity, the PLS regression is applied, presenting values different from 1 in the correlation matrix between the variables and the components (Table 5). Moreover, cross-validation is applied to the components. Equation 9 shows the resulting model to retrieve O_3 during 2014, considering the dataset.

$$O_3 = -0.47\text{PM2.5} - 3.41\text{TMP} - 0.34\text{HUM} - 1371.47\text{B2} + 9449.41\text{B4} - 7852.43\text{B5} - 436.68\text{B7} - 1028.50\text{NDVI} + 4961.14\text{SAVI} + 1178.61\text{EVI} + 66.06 \quad (9)$$

Finally, Eq. 8 allows mapping the O_3 concentration during 2014 (Fig. 8).

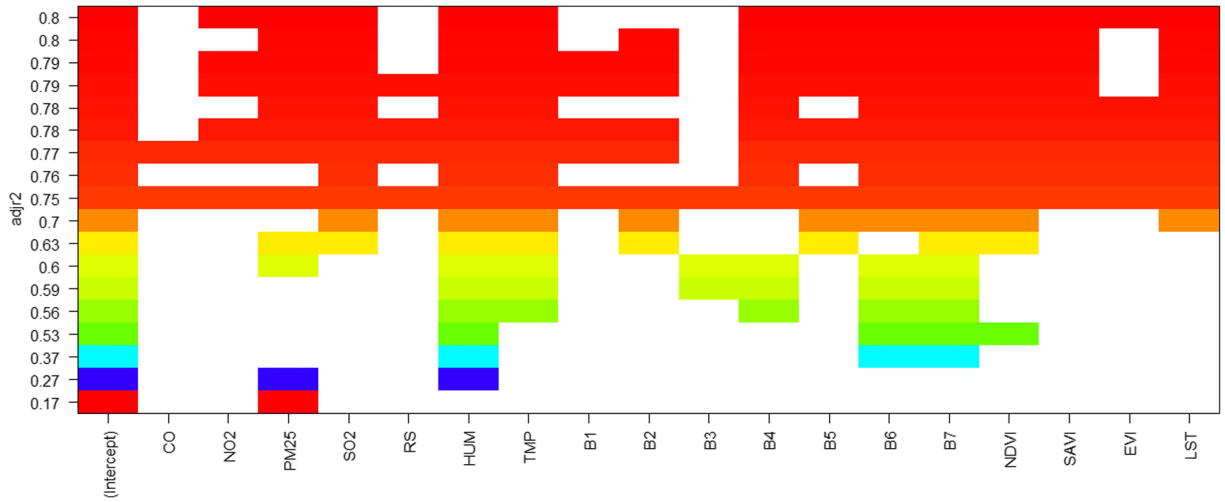


Fig. 5 Variable combinations with their corresponding R^2 values as part of the subset task to select the model with the best fit

Discussion

The main goal of this study was to establish a model to retrieve O_3 from several input variables, implementing a variant of the classical LUR model. In most cases, LUR models are used to model air pollutants from road networks, land use, building density, MODIS AOD, population density and other geographic variables (Ann Becerra et al. 2013; Adam-Poupart et al. 2014; Meng et al. 2016; Wolf et al. 2017; Cattani et al. 2017; Yang et al. 2017). In this study, the variables selected are air pollution measurements, meteorological data (MD) and remote sensing data. The air pollution measurements and MD were obtained from REMMAQ stations.

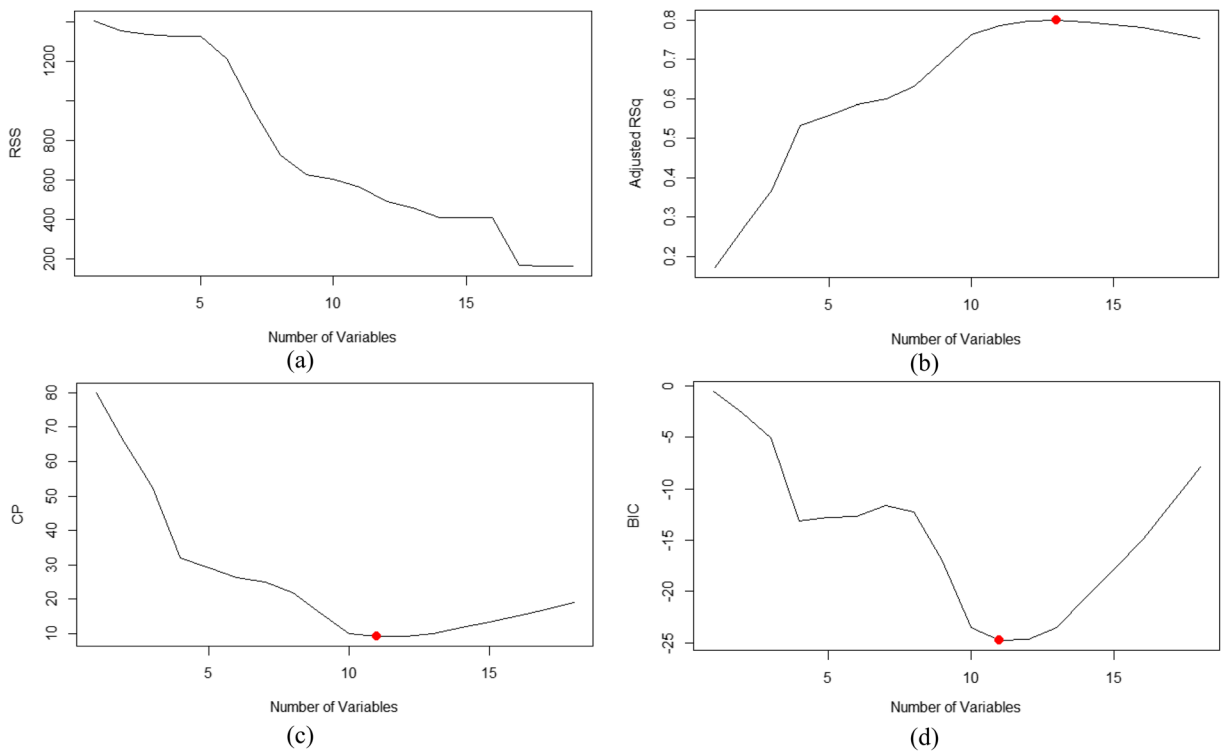


Fig. 6 Subset analysis to select variables with different criteria. **a** RSS, **b** Adj. R^2, **c** CP and d BIC. The red point shows the optimal value of variables for each criterion

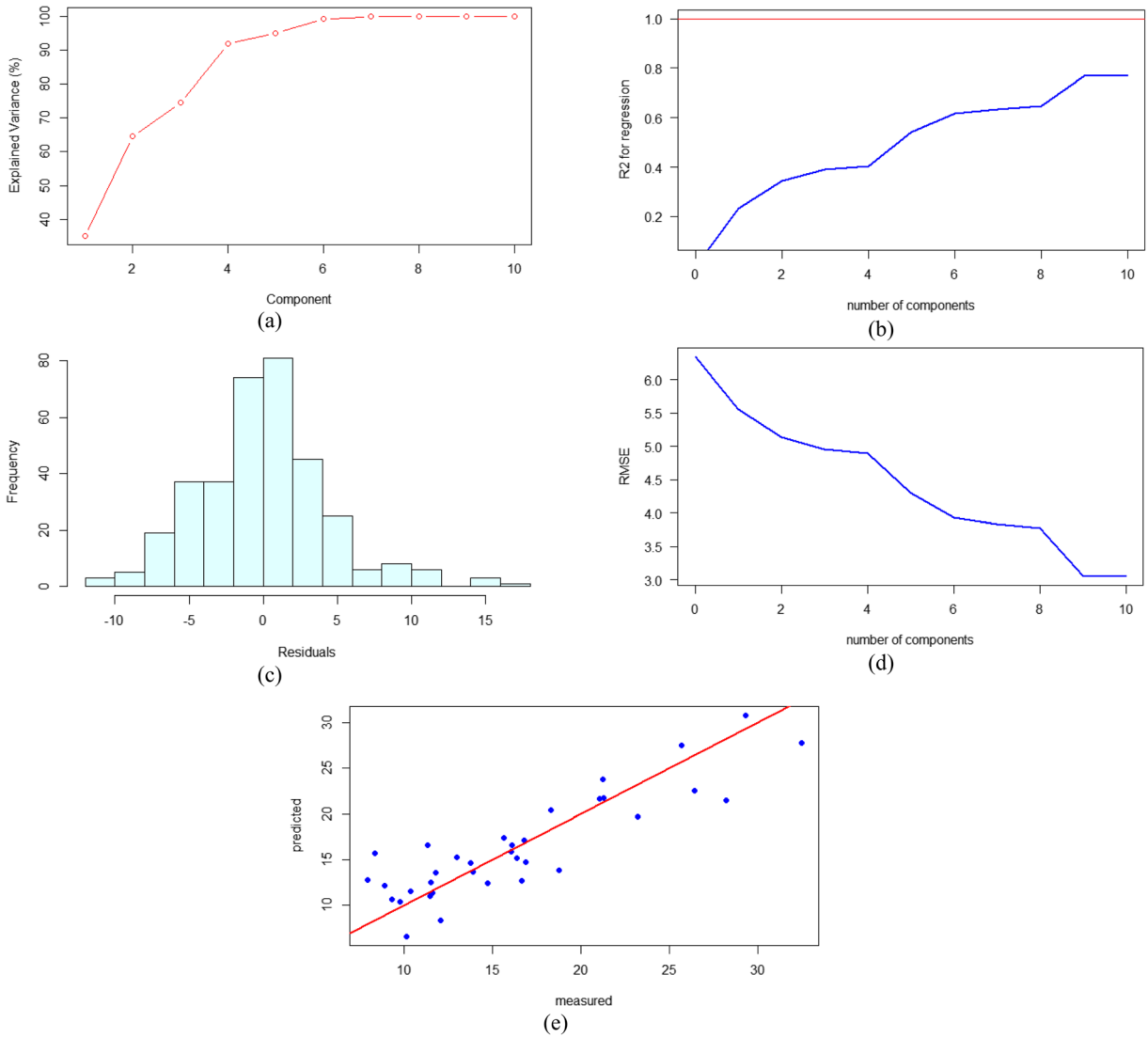


Fig. 7 PLS analysis. **a** The number of components that explain the variance. **b** The number of components to obtain the highest R^2. **c** The histogram of the residuals. **d** The number of components to obtain the lowest RMSE. **e** Measured vs. predicted values with PLS regression

Moreover, considering the accuracy of LUR models in order to retrieve air pollutants (R^2 values between 0.45 and 0.80) (Ann Becerra et al. 2013; Adam-Poupart et al. 2014; Meng et al. 2016; Wolf et al. 2017; Cattani et al. 2017; Yang et al. 2017), ten Landsat 8 images were selected to retrieve O_3 in Quito, Ecuador. Most LUR models use MODIS data. However, MODIS data probably do not have the accuracy and the quality to model pollutants or other environmental variables in middle cities (Teodoro 2015).

To select the predictor variables, a subset was applied considering 19 variables (18 independent variables and O_3 as the dependent variable), obtained a preliminary best fit model with the 18 variables ($R^2 = 0.68$). However, to find the best fit and simplest (with the lowest number of predictors) model, four criteria (RSS, Adj. R^2, CP, BIC) are analysed, resulting in a model with ten independent variables (PM2.5, HUM, TMP, B2, B4, B5, B7, NDVI, SAVI, EVI), showing an R^2 of 0.72 considering stepwise regression. In most of the subsets, the remote sensing data variables B1, B2, B6 and B7 appear, showing the relation between these bands with O_3. Thus, B1 and B2 reflect the blues and violets related to the aerosol presence (Department of the Interior U.S.

Table 4 Variables explained variance by PLS components (t1, t2, …, t6). The red text shows the maximum variance explained with nine components, considering O_3 as the dependent variable

Variable	t1	t2	t3	t4	t5	t6	t7	t8	t9
PM2.5	0.148	0.655	0.660	0.787	0.897	0.999	1.000	1.000	1.000
HUM	0.212	0.433	0.442	0.593	0.775	1.000	1.000	1.000	1.000
TMP	0.017	0.350	0.902	0.978	0.979	1.000	1.000	1.000	1.000
B2	0.611	0.918	0.918	0.947	0.955	0.966	0.998	1.000	1.000
B4	0.609	0.934	0.948	0.994	0.995	0.998	0.998	1.000	1.000
B5	0.123	0.158	0.362	0.994	0.997	0.999	1.000	1.000	1.000
B7	0.460	0.714	0.777	0.951	0.952	0.974	0.996	1.000	1.000
NDVI	0.515	0.873	0.904	0.987	0.987	0.993	0.994	1.000	1.000
SAVI	0.435	0.740	0.805	0.989	0.990	1.000	1.000	1.000	1.000
EVI	0.387	0.677	0.729	0.957	0.958	0.991	0.999	1.000	1.000
R^2	0.232	0.345	0.390	0.404	0.541	0.617	0.634	0.646	0.768

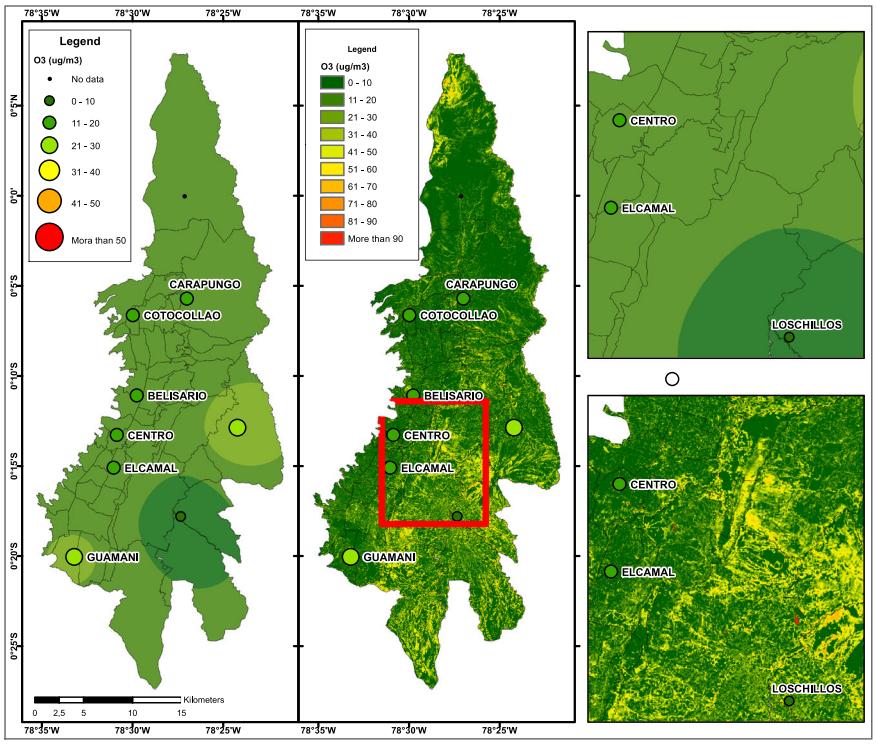
Geological Survey 2016). Additionally, B6 and B7 reflect the short infrared related to greenhouse gas absorption (North 2015). Some studies that use LUR models employed stepwise regression to automatically find the predictors in a model (Ayres-Sampaio et al. 2014; Olmanson et al. 2016). However, the main problem with stepwise regression is not allowing a multicollinearity analysis (NCSS and LLC). PLS regression is used in some studies to compute the LUR model (Adam-Poupart et al. 2014; Wang et al. 2016) to avoid multicollinearity. PLS builds a model with latent variables (components) as independent variables (Williams et al. 2013). Moreover, PLS regression is used when we have a model with few observations (Chi et al. 2018). If a high correlation is present between variables, a PLS regression is used to build the model, where nine components explain most of the variance and obtained an R^2 value of 0.768. This value is higher than R^2 in the stepwise regression ($R^2 = 0.72$) and avoids the multicollinearity of remote sensing variables.

The final model can be mapped, in comparison with other techniques, such as thematic point maps, interpolation or geostatistical analysis (Fig. 8), showing a robust perception of spatial concentration of O_3 in the city, and these maps can be used as input to make a more accurate air pollution analysis.

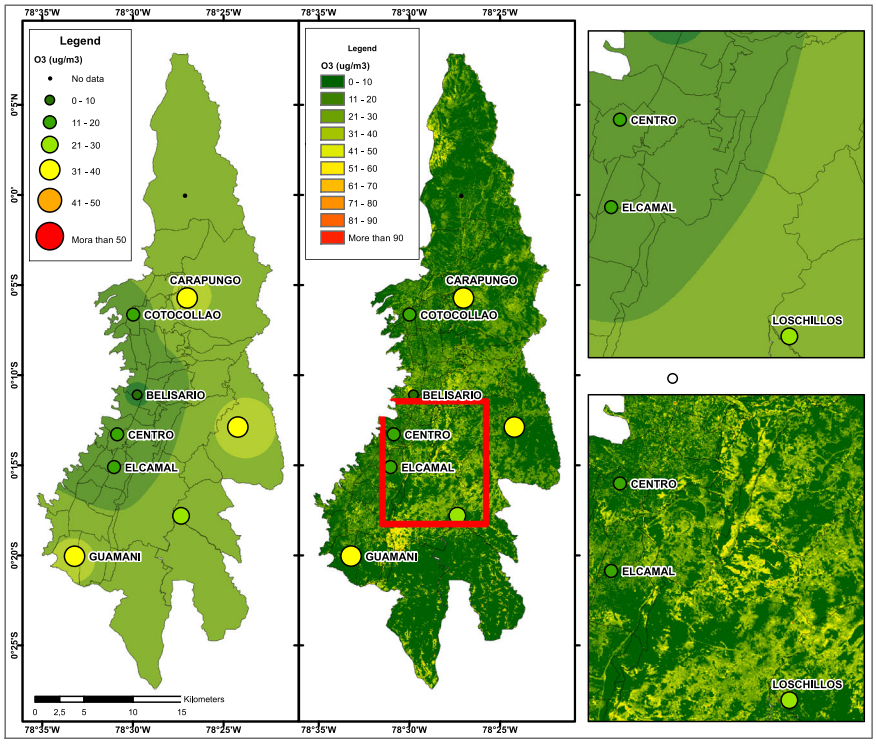
The limitation is the few observations used to build the model because our model requires some data from the REMMAQ stations, and sometimes, these data are incomplete or unavailable. On the other

Table 5 Correlation matrix between the variables and the PLS components

Variable	t1	t2	t3	t4	t5	t6	t7	t8	t9
PM2.5	− 0.38514	− 0.71215	− 0.06697	− 0.35607	0.33247	− 0.31854	0.03311	− 0.01169	0.00006
HUM	− 0.46074	− 0.46963	− 0.09577	0.38916	− 0.42616	0.47404	− 0.01179	0.00889	− 0.00005
TMP	0.13159	0.57670	− 0.74288	− 0.27615	− 0.02596	0.14489	0.01716	0.00092	− 0.00003
B2	0.78187	− 0.55363	0.01333	0.17146	− 0.08946	− 0.10447	0.17864	− 0.04175	− 0.00072
B4	0.78063	− 0.56987	− 0.11695	0.21595	− 0.01183	0.05406	0.00119	− 0.04905	0.00434
B5	0.35068	− 0.18755	− 0.45108	0.79548	0.05186	− 0.04768	0.01853	− 0.01613	− 0.00254
B7	0.67796	− 0.50463	− 0.25075	0.41702	0.02133	− 0.14883	− 0.14855	0.06528	− 0.00017
NDVI	− 0.71749	0.59850	− 0.17601	0.28787	− 0.02397	− 0.07591	− 0.02313	− 0.07924	− 0.00011
SAVI	− 0.65920	0.55261	− 0.25518	0.42854	0.03350	− 0.09995	0.00418	0.00529	0.00143
EVI	− 0.62209	0.53862	− 0.22889	0.47707	0.02165	− 0.18231	0.08867	0.03466	0.00153
O_3	0.48204	0.33513	0.21357	0.11803	0.36972	0.27520	0.13251	0.10736	0.34908



(a)



(b)

Fig. 8 Maps of O_3 during 2014. **a** January and **b** July maps obtained from Eq. 8. The left map is with an inverse distance weighting (IDW) technique while the centre map is applying the O_3 model in all the study area. The right maps are a zoom in an assessment area (red square)

hand, the remote sensing variables depend on the number of clouds. Quito is known as a city with a high cloud density during the year (Alvarez-Mendoza et al. 2018a), and this factor limits the computation of LUR models. A possible alternative can be to combine different sensors with high spatial and temporal resolution and use similar techniques to PLS to compute the model.

Another limitation is the generation of a raster to each independent variable. In the case of remote sensing, data are not a problem considering all images over the study area, but the air pollutant measurements and MD raster can be limited. They were obtained with a geostatistical technique as inverse distance weighting (IDW) (de Mesnard 2013). Nevertheless, this kind of technique works fine in a region with some stations, but in Quito, we only have nine stations (Fig. 8). Therefore, in future work, we will propose the use of only remote sensing data to spatialize air pollutants in Quito.

Conclusion

A spatial estimation was performed in Quito to obtain the O_3 spatial concentration in 2014. The spatial estimation was computed by a variant of LUR models with PLS regression. LUR models can explain the spatial concentration of an air pollutant, helping in urban planning, environmental analysis and governmental decisions. Moreover, the idea of having a variant of LUR models with variables from remote sensing sensors different from MODIS will help to build more accurate models. The main limitation is related to the small quantities of field data available. In future work, we will try to find new alternatives only considering the use of remote sensing data as input without other field data variables.

Acknowledgements This study is part of a PhD thesis in Surveying Engineering at the University of Porto, Portugal, supported by the Salesian Polytechnic University, Ecuador. This work was supervised at the University of Porto by Prof. Ana Cláudia Teodoro. The statistical analysis and regression models were supervised by Prof. Lenin Ramirez. We thank to Kathy Copo and Michelle Burgos for assistance with data acquisition to evaluate some initial parts of paper.

References

Adam-Poupart, A., Brand, A., Fournier, M., Jerrett, M., & Smargiassi, A. (2014). Spatiotemporal modeling of ozone levels in Quebec (Canada): a comparison of kriging, land-use regression (LUR), and combined Bayesian maximum entropy–LUR approaches. *Environmental Health Perspectives, 122*, 970–976. https://doi.org/10.1289/ehp.1306566.

Alvarez, C. I., Padilla Almeida, O. (2016). Estimación de la contaminación del aire por PM10 en Quito a través de índices ambientales con imágenes LANDSAT ETM+. Revista Cartográfica 135–147.

Alvarez-Mendoza, C. I., Teodoro, A., & Ramirez-Cando, L. (2018a). Improving NDVI by removing cirrus clouds with optical remote sensing data from Landsat-8 – a case study in Quito, Ecuador. *Remote Sensing Applications: Society and Environment, 13*, 257–274. https://doi.org/10.1016/j.rsase.2018.11.008.

Alvarez-Mendoza, C. I., Teodoro, A., Torres, N., et al. (2018b). Comparison of satellite remote sensing data in the retrieve of PM10 air pollutant over Quito, Ecuador. In: Thilo Erbertseder, Nektarios Chrysoulakis YZ (ed) Remote Sensing Technologies and Applications in Urban Environments III. Berlin, p 107930I–10793–12.

Ann Becerra, T., Wilhelm, M., Olsen, J., et al. (2013). Ambient air pollution and autism in Los Angeles County, California. *Environmental Health Perspectives, 121*, 380–386. https://doi.org/10.1289/ehp.1205827.

As-syakur, A. R., Adnyana, I. W. S., Arthana, I. W., & Nuarsa, I. W. (2012). Enhanced built-up and bareness index (EBBI) for mapping built-up and bare land in an urban area. *Remote Sensing, 4*, 2957–2970. https://doi.org/10.3390/rs4102957.

Ayres-Sampaio, D., Teodoro, A. C., Sillero, N., Santos, C., Fonseca, J., & Freitas, A. (2014). An investigation of the environmental determinants of asthma hospitalizations: an applied spatial approach. *Applied Geography, 47*, 10–19. https://doi.org/10.1016/j.apgeog.2013.11.011.

Bilal, M., Nichol, J. E., Bleiweiss, M. P., & Dubois, D. (2013). A simplified high resolution MODIS aerosol retrieval algorithm (SARA) for use over mixed surfaces. *Remote Sensing of Environment, 136*, 135–145. https://doi.org/10.1016/j.rse.2013.04.014.

Braun, D., Damm, A., Hein, L., Petchey, O. L., & Schaepman, M. E. (2018). Spatio-temporal trends and trade-offs in ecosystem services: an Earth observation based assessment for Switzerland between 2004 and 2014. *Ecological Indicators, 89*, 828–839. https://doi.org/10.1016/J.ECOLIND.2017.10.016.

Cattani, G., Gaeta, A., Di Menno di Bucchianico, A., di Menno di Bucchianico, A., de Santis, A., Gaddi, R., Cusano, M., Ancona, C., Badaloni, C., Forastiere, F., Gariazzo, C., Sozzi, R., Inglessis, M., Silibello, C., Salvatori, E., Manes, F., & Cesaroni, G. (2017). Development of land-use regression models for exposure assessment to ultrafine particles in Rome, Italy. *Atmospheric Environment, 156*, 52–60. https://doi.org/10.1016/J.ATMOSENV.2017.02.028.

Cazorla, M. (2016). Air quality over a populated andean region: Insights from measurements of ozone, NO, and boundary layer depths. *Atmospheric Pollution Research, 7*, 66–74. https://doi.org/10.1016/j.apr.2015.07.006.

Chen, G., & Meentemeyer, R. (2016). Remote sensing of forest damage by diseases and insects. In Q. Weng (Ed.), *Remote Sensing for Sustainability* (2017th ed., p. 357). Boca Raton: CRC Press.

Chen, L., Bai, Z., Kong, S., Han, B., You, Y., Ding, X., du, S., & Liu, A. (2010). A land use regression for predicting NO2 and PM10 concentrations in different seasons in Tianjin region, China. *Journal of Environmental Sciences, 22*, 1364–1373. https://doi.org/10.1016/S1001-0742(09)60263-1.

Chen, Y., Han, W., Chen, S., & Tong, L. (2014). Estimating ground-level PM2.5 concentration using Landsat 8 in Chengdu, China. *Proceedings of SPIE, 9259*, 925917–925931. https://doi.org/10.1117/12.2068886.

Chi, Y., Shi, H., Zheng, W., & Sun, J. (2018). Simulating spatial distribution of coastal soil carbon content using a comprehensive land surface factor system based on remote sensing. *Science of the Total Environment, 628–629*, 384–399. https://doi.org/10.1016/j.scitotenv.2018.02.052.

Daac, N. L. P, Falls, S., March, S. D. (2012). MODIS land products quality assurance tutorial: Part-1 how to find , understand , and use the quality assurance information for MODIS land products. 1–15.

de Mesnard, L. (2013). Pollution models and inverse distance weighting: some critical remarks. *Computational Geosciences, 52*, 459–469. https://doi.org/10.1016/J.CAGEO.2012.11.002.

Department of the Interior U.S. Geological Survey (2016). Landsat 8 (L8) Data Users Handbook http://www.crisp.nus.edu.sg/~research/tutorial/optical.htm. Accessed 21 Jul 2017.

Gholizadeh, H., & Robeson, S. M. (2016). Revisiting empirical ocean-colour algorithms for remote estimation of chlorophyll-a content on a global scale revisiting empirical ocean-colour algorithms for remote estimation of chlorophyll-a content on a global scale. *International Journal of Remote Sensing, 37*, 2682–2705. https://doi.org/10.1080/01431161.2016.1183834org/10.1080/01431161.2016.1183834.

Gillespie, A. (2014). Land surface emissivity. In E. Njoku (Ed.), *Encyclopedia of remote sensing* (p. 939). New York: Springer-Verlag.

Habermann, M., Billger, M., & Haeger-Eugensson, M. (2015). Land use regression as method to model air pollution. Previous results for Gothenburg/Sweden. *Procedia Engineering, 115*, 21–28. https://doi.org/10.1016/J.PROENG.2015.07.350.

Instituto Nacional de Meteorología e Hidrología (2016). Boletín Climatológico Anual 2015.

Jia, K., Liang, S., Zhang, L., Wei, X., Yao, Y., & Xie, X. (2014). Forest cover classification using Landsat ETM+ data and time series MODIS NDVI data. *International Journal of Applied Earth Observation and Geoinformation, 33*, 32–38. https://doi.org/10.1016/j.jag.2014.04.015.

Larkin, A., Geddes, J. A., Martin, R. V., Xiao, Q., Liu, Y., Marshall, J. D., Brauer, M., & Hystad, P. (2017). Global land use regression model for nitrogen dioxide air pollution. *Environmental Science & Technology, 51*, 6957–6964. https://doi.org/10.1021/acs.est.7b01148.

Lee, P., Saylor, R., & McQueen, J. (2018). Air quality monitoring and forecasting. *Atmosphere (Basel), 9*, 89. https://doi.org/10.3390/atmos9030089.

Liang, C.-S., Liu, H., He, K.-B., Ma, Y.-L. (2016). Assessment of regional air quality by a concentration-dependent Pollution Permeation Index OPEN. doi: https://doi.org/10.1038/srep34891.

Liew, S. C. (2001). Principles of remote sensing. In: Cent. Remote Imaging, Sens. Process. Cris.

Liu, Y., Franklin, M., Kahn, R., & Koutrakis, P. (2007). Using aerosol optical thickness to predict ground-level PM2.5 concentrations in the St. Louis area: a comparison between MISR and MODIS. *Remote Sensing of Environment, 107*, 33–44. https://doi.org/10.1016/j.rse.2006.05.022.

Meng, X., Chen, L., Cai, J., Zou, B., Wu, C. F., Fu, Q., Zhang, Y., Liu, Y., & Kan, H. (2015). A land use regression model for estimating the NO2 concentration in Shanghai, China. *Environmental Research, 137*, 308–315. https://doi.org/10.1016/j.envres.2015.01.003.

Meng, X., Fu, Q., Ma, Z., Chen, L., Zou, B., Zhang, Y., Xue, W., Wang, J., Wang, D., Kan, H., & Liu, Y. (2016). Estimating ground-level PM10 in a Chinese city by combining satellite data, meteorological information and a land use regression model. *Environmental Pollution, 208*, 177–184. https://doi.org/10.1016/J.ENVPOL.2015.09.042.

Mok, K. M., Yuen, K. V., Hoi, K. I., Chao, K. M., & Lopes, D. (2018). Predicting ground-level ozone concentrations by adaptive Bayesian model averaging of statistical seasonal models. *Stochastic Environmental Research and Risk Assessment, 32*, 1283–1297. https://doi.org/10.1007/s00477-017-1473-1.

Monks, P. S., Archibald, A. T., Colette, A., Cooper, O., Coyle, M., Derwent, R., Fowler, D., Granier, C., Law, K. S., Mills, G. E., Stevenson, D. S., Tarasova, O., Thouret, V., von Schneidemesser, E., Sommariva, R., Wild, O., & Williams, M. L. (2015). Tropospheric ozone and its precursors from the urban to the global scale from air quality to short-lived climate forcer. *Atmospheric Chemistry and Physics, 15*, 8889–8973. https://doi.org/10.5194/acp-15-8889-2015.

NASA EOSDIS (2018). Remote Sensors In: Earthdata. https://earthdata.nasa.gov/user-resources/remote-sensors. Accessed 21 Jul 2018 NCSS, LLC Stepwise Regression.

North, G. R. (2015). Climate and climate change - Greenhouse effect. *Encyclopedia of Atmospheric Sciences*, 80–86. https://doi.org/10.1016/B978-0-12-382225-3.00470-9.

Olmanson, L. G., Brezonik, P. L., Finlay, J. C., & Bauer, M. E. (2016). Comparison of Landsat 8 and Landsat 7 for regional measurements of CDOM and water clarity in lakes. Remote Sensing of Environment, 185, 119–128. https://doi.org/10.1016/j.rse.2016.01.007.

Secretaria del Ambiente de Quito (2018). Red Metropolitana de Monitoreo Atmosférico de Quito. http://www.quitoambiente.gob.ec/ambiente/index.php/generalidades. Accessed 26 Jun 2018.

Sicard, P., Anav, A., De Marco, A., & Paoletti, E. (2017). Projected global ground-level ozone impacts on vegetation under different emission and climate scenarios. *Atmospheric Chemistry and Physics, 17*, 12177–12196. https://doi.org/10.5194/acp-17-12177-2017.

Sobrino, J. A., Jiménez-Muñoz, J. C., Sòria, G., et al. (2008). Land surface emissivity retrieval from different VNIR and TIR sensors. *IEEE Transactions on Geoscience and Remote Sensing, 46*, 316–327. https://doi.org/10.1109/TGRS.2007.904834.

Stafoggia, M., Schwartz, J., Badaloni, C., Bellander, T., Alessandrini, E., Cattani, G., de' Donato, F., Gaeta, A., Leone, G., Lyapustin, A., Sorek-Hamer, M., de Hoogh, K., di, Q., Forastiere, F., & Kloog, I. (2017). Estimation of daily PM10 concentrations in Italy (2006–2012) using finely resolved satellite data, land use variables and meteorology. *Environment International, 99*, 234–244. https://doi.org/10.1016/J.ENVINT.2016.11.024.

Teodoro, A. (2015). *A study on the quality of the vegetation index obtainded from MODIS data* (pp. 3365–3368). Yokohama: IGARSS.

U.S. Geological Survey (2016). Product Guide: Landsat 8 Surface Reflectance Product

US Department of Commerce (2018). NOAA ESRL ESRL Global Monitoring Division - Ozone and Water Vapor Group. https://www.esrl.noaa.gov/gmd/ozwv/surfoz/. Accessed 23 May 2018.

US EPA (2014). Report of the Environment: Ozone Concentrations

USGS (2017). Product guide: Landsat surface reflectance-derived spectral indices

Vermote, E., Justice, C., Claverie, M., & Franch, B. (2016). Preliminary analysis of the performance of the Landsat 8/OLI land surface reflectance product. *Remote Sensing of Environment, 185*, 46–56. https://doi.org/10.1016/J.RSE.2016.04.008.

Vieira, D., Teodoro, A., & Gomes, A. (2016). Analysing land surface temperature variations during Fogo Island (Cape Verde) 2014-2015 eruption with Landsat 8 images. *Proceedings of SPIE, 10005*, 1000508. https://doi.org/10.1117/12.2241382.

Wang, M., Sampson, P. D., Hu, J., Kleeman, M., Keller, J. P., Olives, C., Szpiro, A. A., Vedal, S., & Kaufman, J. D. (2016). Combining land-use regression and chemical transport modeling in a spatiotemporal geostatistical model for ozone and PM 2.5. *Environmental Science & Technology, 50*, 5111–5118. https://doi.org/10.1021/acs.est.5b06001.

WHO (World Health Organization). (2013). Health risks of ozone from long-range transboundary air pollution. *Journal of Chemical Information and Modeling, 53*, 1689–1699. https://doi.org/10.1017/CBO9781107415324.004.

Williams, L. J., Abdi, H., & Williams, L. J. (2013). Partial least squares methods: Partial least squares correlation and partial Least Square regression. In B. Reisfeld & A. N. Mayeno (Eds.), *Computational toxicology* (Vol. II, pp. 549–579). Totowa: Humana Press.

Wolf, K., Cyrys, J., Harciníková, T., Gu, J., Kusch, T., Hampel, R., Schneider, A., & Peters, A. (2017). Land use regression modeling of ultrafine particles, ozone, nitrogen oxides and markers of particulate matter pollution in Augsburg, Germany. *Science of the Total Environment, 579*, 1531–1540. https://doi.org/10.1016/J.SCITOTENV.2016.11.160.

Yang, X., Zheng, Y., Geng, G., Liu, H., Man, H., Lv, Z., He, K., & de Hoogh, K. (2017). Development of PM2.5 and NO2 models in a LUR framework incorporating satellite remote sensing and air quality model data in Pearl River Delta region, China. *Environmental Pollution, 226*, 143–153. https://doi.org/10.1016/J.ENVPOL.2017.03.079.

Zhang, J., Hu, J., Lian, J., Fan, Z., Ouyang, X., & Ye, W. (2016). Seeing the forest from drones: testing the potential of lightweight drones as a tool for long-term forest monitoring. *Biological Conservation, 198*, 60–69. https://doi.org/10.1016/j.biocon.2016.03.027.

Zhang, X., Chu, Y., Wang, Y., & Zhang, K. (2018). Predicting daily PM2.5 concentrations in Texas using high-resolution satellite aerosol optical depth. *Science of the Total Environment, 631–632*, 904–911. https://doi.org/10.1016/J.SCITOTENV.2018.02.255.

Zheng, S., Zhou, X., Singh, R., Wu, Y., Ye, Y., & Wu, C. (2017). The spatiotemporal distribution of air pollutants and their relationship with land-use patterns in Hangzhou city, China. *Atmosphere (Basel), 8*, 110. https://doi.org/10.3390/atmos8060110.

Annex II

Evaluation of automatic cloud removal method for high elevation areas in Landsat 8 OLI images to improve environmental indexes computation

César I. Alvarez, Ana Teodoro, Alfonso Tierra

César I. Alvarez, Ana Teodoro, Alfonso Tierra, "Evaluation of automatic cloud removal method for high elevation areas in Landsat 8 OLI images to improve environmental indexes computation," Proc. SPIE 10428, Earth Resources and Environmental Remote Sensing/GIS Applications VIII, 1042809 (5 October 2017); doi: 10.1117/12.2277844

SPIE.

Event: SPIE Remote Sensing, 2017, Warsaw, Poland

Evaluation of Automatic Cloud Removal Method for high elevation areas in Landsat 8 OLI images to improve environmental indexes computation

César I. Alvarez*[a,b], Ana Teodoro[a,c], Alfonso Tierra[d]
[a]Dep.of Geosciences, Environment and Land Planning, Faculty of Sciences, University of Porto, 4169-007 Porto, Portugal; [b] Grupo GIADES, Carrera de Ingeniería Ambiental, Universidad Politécnica Salesiana, Quito, Ecuador; [c]Earth Sciences Institute (ICT), Pole of the FCUP, University of Porto, Portugal; [d]Grupo de Inv. Geoespacial. U. Fuerzas Armadas ESPE. Av. Gral Rumiñahui s/n. Sangolquí, Ecuador. P.O.Box 171-5-231B.

ABSTRACT

Thin clouds in the optical remote sensing data are frequent and in most of the cases don't allow to have a pure surface data in order to calculate some indexes as Normalized Difference Vegetation Index (NDVI). This paper aims to evaluate the Automatic Cloud Removal Method (ACRM) algorithm over a high elevation city like Quito (Ecuador), with an altitude of 2800 meters above sea level, where the clouds are presented all the year. The ACRM is an algorithm that considers a linear regression between each Landsat 8 OLI band and the Cirrus band using the slope obtained with the linear regression established. This algorithm was employed without any reference image or mask to try to remove the clouds. The results of the application of the ACRM algorithm over Quito didn't show a good performance. Therefore, was considered improving this algorithm using a different slope value data (ACMR Improved). After, the NDVI computation was compared with a reference NDVI MODIS data (MOD13Q1). The ACMR Improved algorithm had a successful result when compared with the original ACRM algorithm. In the future, this Improved ACRM algorithm needs to be tested in different regions of the world with different conditions to evaluate if the algorithm works successfully for all conditions.

Keywords: Remove Cloud, Landsat, NDVI, Cirrus Band, Quito

1. INTRODUCTION

One of the principal problems that are considered in optical remote sensing is the cloud density over some areas of the world[1], understanding that in some areas like South America[2] and places with high mountains like Andean Region[3] the presence of high cloud density is real during most of the year, discussing if the remote sensing data is real useful to calculate some environmental parameters as Normalized Difference Vegetation Index (NDVI)[4]. In the studies previous referred[2], Landsat images are considered. Landsat is a land optical remote sensing program that for four decades provides images that could be used in different areas, as agriculture, geology, forestry, environment and mapping[5]. The last satellite of this program is Landsat 8, which include two sensors: (1) Operational Land Imager (OLI) and; (2) Thermal Infrared Sensor (TIR). Moreover, Landsat 8 OLI provides detection of high-altitude cloud contamination that may not be detectable in other spectral bands[6].

Some algorithms had been developed with the challenge to try to remove thin clouds in different regions considering Landsat 8 OLI imagery. Nevertheless, these algorithms use a Landsat reference image from other dates to patch the cloud area[7–9]. Other algorithms combine Landsat with other sensors[10] and others use the same image considering the Cirrus band (B9) in Landsat 8 in order to remove thin clouds[11–13].

The main idea in this work was to evaluate and improve for the Andean Region (Quito, Ecuador) one of the algorithms developed to remove thin clouds called Automatic Cloud Removal Method (ACRM)[13]. This algorithm was originally evaluated in Sidney, Australia, which have conditions very different from Quito. The ACRM algorithm established a linear regression between each band in the OLI sensor with B9, considering some selected areas in the image[13]. The

Earth Resources and Environmental Remote Sensing/GIS Applications VIII, edited by Ulrich Michel, Karsten Schulz
Proc. of SPIE Vol. 10428, 1042809 · © 2017 SPIE · CCC code: 0277-786X/17/$18 · doi: 10.1117/12.2277844

concept is to find the best fit area with the highest coefficient of determination (R^2) to generalize for the entire image the application of the algorithm to remove thin clouds. The algorithm in the original study was applied with success[13], nevertheless, in different areas, as Quito, the results obtained are not satisfactory, let us to consider that the algorithm should be improved. The idea is to find the best fit considering a good R^2 and the correct slope (α) for the study area.

2. STUDY AREA AND DATA

Quito is the capital of Ecuador with an elevation area with approximately 2800 meters above sea level. Also, the city is in the center of Andean Region and it is influenced by the equatorial line with latitudes nearest to 0º and being in a Tropical Region (Figure 1). Moreover, Quito doesn't present clearly seasons. The mean temperature during the year has a mean in minimum about 9.0°C and a maximum 25.4 °C, also presented a high precipitation near to 1126 mm on 2015 that let to have a high-density cloud every year[14].

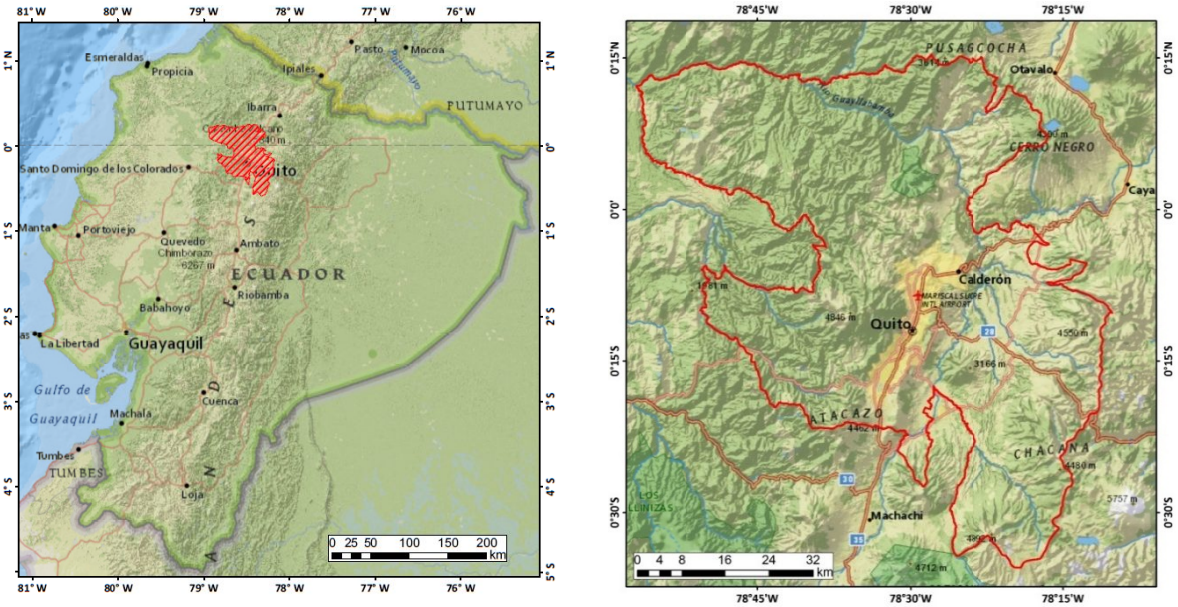


Figure 1. Study area location (Quito, Ecuador).

The images considered to this study were six Landsat 8 OLI L1T scenes in different dates. Four images (Figure 2) from study area (Quito Path:10 Row:60) and two images to compare and to evaluate the original algorithm: one from Pedernales, Ecuador (Path:11; Row:60) and other from Sidney, Australia (Path:89; Row:83). Sidney was the area considered in the original algorithm[13]. Additionally to these images were also considered the MODIS MOD13Q1 product (tiles H10V08 and H10V09) for the study area in order to compare with NDVI calculated with ACRM algorithm (more details in Section 3.3).

3. METHODOLOGY

Landsat 8 is the last satellite in orbit from Landsat Data Continuity Mission (LDCM). It was launched in February, 2013 and it has two push-broom instruments: the Operational Land Image (OLI) with 30 meters of spatial resolution and the Thermal Infrared Sensor (TIRS) with 100 meters of spatial resolution resampled to 30 meters[5]. The two sensors over Landsat 8 have 11 spectral bands where the principal attention considering this work is the Band 9 - the Cirrus band[5]. In this work, we used Landsat 8 images to try to remove thin clouds using Cirrus Band (B9) and considering the Automatic Cloud Remove Method (ACRM)[13]. All the processing steps were implemented in R programming language[15] and association packages raster version 2.5-8[16] to work with raster images, rgdal version 1.1[17] and gdalutilities version 2.0.1.7[18] to work with geospatial data.

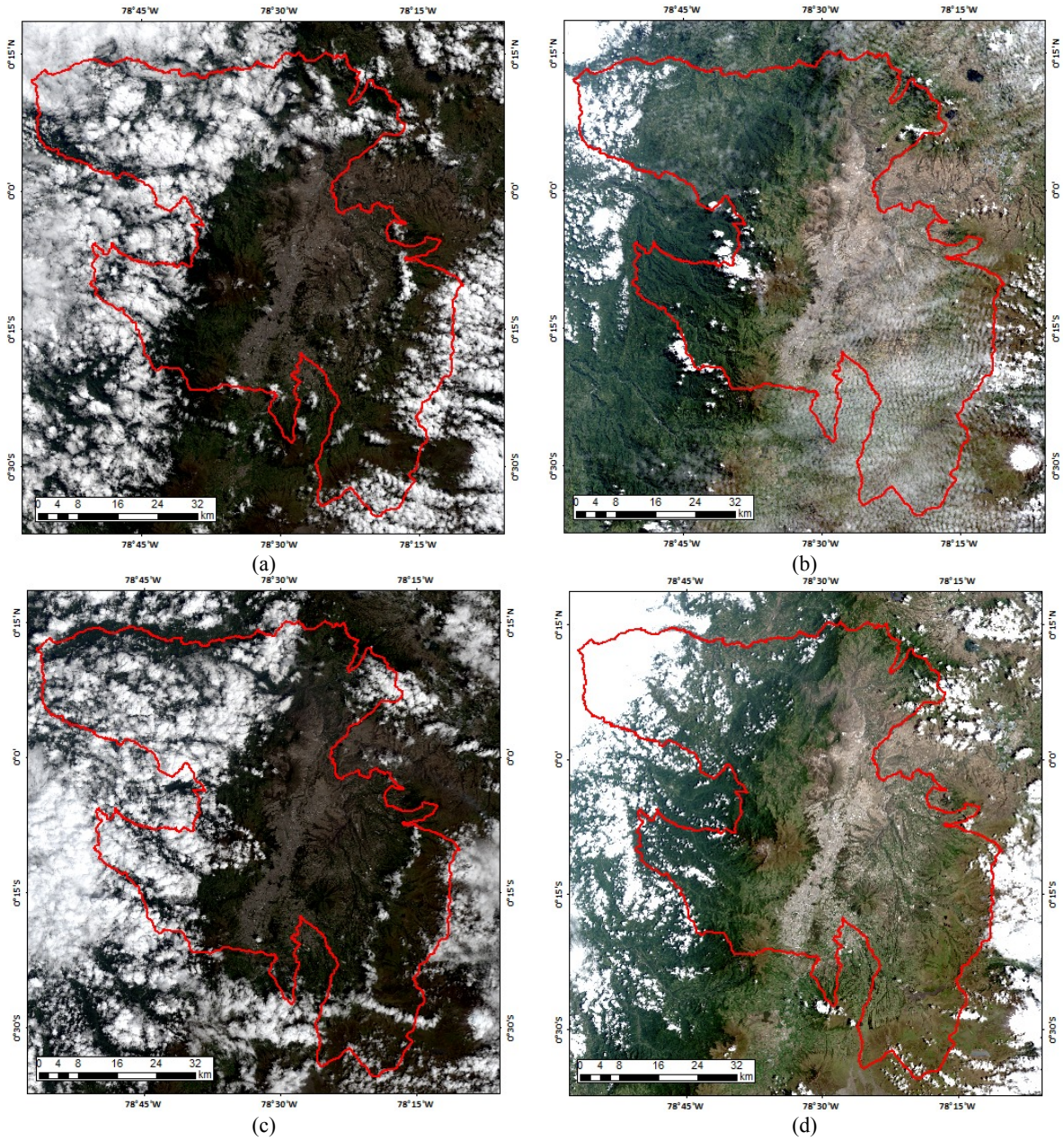


Figure 2. Quito Landsat 8 OLI: (a) Image from 2013/07/07; (b) Image from 2014/07/26; (c) Image from 2015/08/30; (d) Image from 2016/10/19.

3.1 Automatic Cloud Removal Method (ACRM)

The Cirrus band provides a way to remove thin clouds in an image considering that the noise and clouds are part of the original image, in each band. Considering this, the Equation (1) explains how the algorithm is computed.

$$DN(u,v) = x_i^f(u,v) + x_i^c(u,v) \qquad i = 1,2,3,4,5,6,7. \tag{1}$$

Where *DN(u,v)* represents the Digital Number in each band, $x_i^f(u,v)$ represents the pure surface pixel data from each Landsat OLI band, and $x_i^c(u,v)$ represents the pixel with noise (clouds). Consequently, the challenge is to obtain $x_i^c(u,v)$ represented as a linear relation to pixel data in the Cirrus band $c(u,v)$, as showed in Equation (2).

$$x_i^c(u,v) = \alpha_i\left[c(u,v) - \min\{c(u,v)\}\right] \tag{2}$$

Replacing the Equation (2), in Equation (1) is obtained the Equation (3).

$$x_i^f(u,v) = DN(u,v) - \alpha_i\left[c(u,v) - \min\{c(u,v)\}\right] \tag{3}$$

In Equation (3) is showed the final pure surface pixel data $x_i^f(u,v)$. The objective is to obtain the slope α_i in a homogenous area and established a linear regression between each OLI Band (B1-B7) and B9. Accordingly, the challenge is to obtain this homogeneous area considering that it is a part of the entire image. For this, two possibilities should to be analyzed. Firstly, the homogeneous area can be determined manually considering aspects as the water areas have strong absorption in the NIR and MIR bands nearest to zero and the contaminated pixel in these water areas can show the clouds presence, in this case can be considered. The method is effective, but it doesn't have a good accuracy considering other physics aspects in water zones like waves dynamics. On the other hand, the automatic homogenous area identification can be used considering some aspects like different soil cover areas or vegetation areas. In this work, the second approach (automatic) was chosen in order to obtain the slope value because the study area doesn't present water bodies to be considered, generating regular zones of 10x10 km which cover all the study area (Figure 3). A total of 90 zones were tested, divided around the entire image.

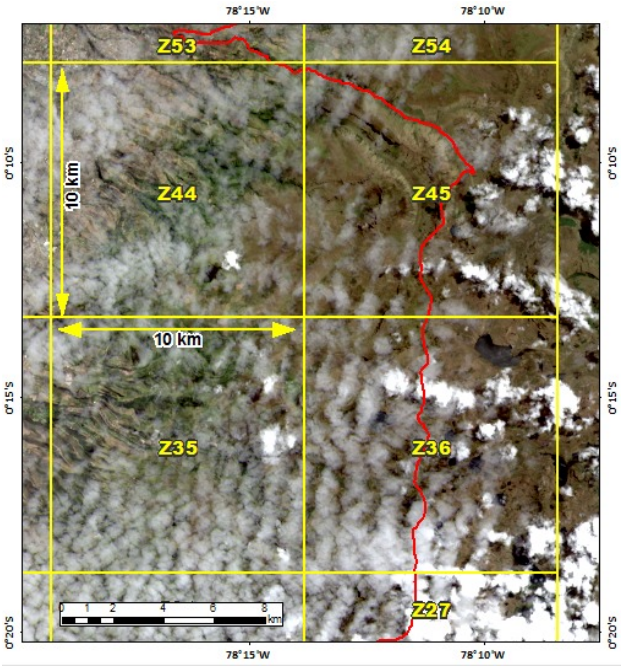


Figure 3. Example of some areas evaluated to test the algorithm over Quito. Each area has 10x10 km.

3.2 Normalized Difference Vegetation Index (NDVI)

The NDVI is one of the most used remote sensing indices[19,20]. It allows to obtain information about the greenest vegetation considering Red and NIR bands[21], as shown in Equation (4).

$$NDVI = \frac{NIR - RED}{NIR + RED} \tag{4}$$

NDVI also could be considered an environmental index[20], besides that it has a great relation with Land Surface[22] due to their relationship to other variables like temperature, vegetation, humidity, etc.

NDVI was calculated considering Landsat 8 image from Quito (Figure 4) and applying the ACRM algorithm (Figure 5), considering Red Band B4 (0.636 - 0.673 µm) and NIR Band B5 (0.851 - 0.879 µm), as given in Equation (4).

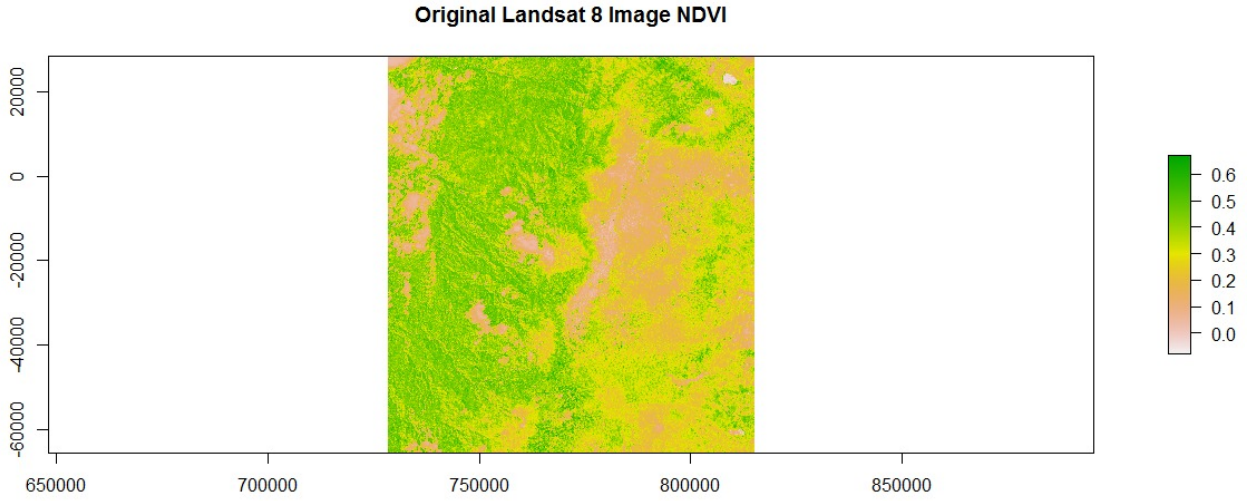


Figure 4. NDVI computed in Landsat 8 for the study area.

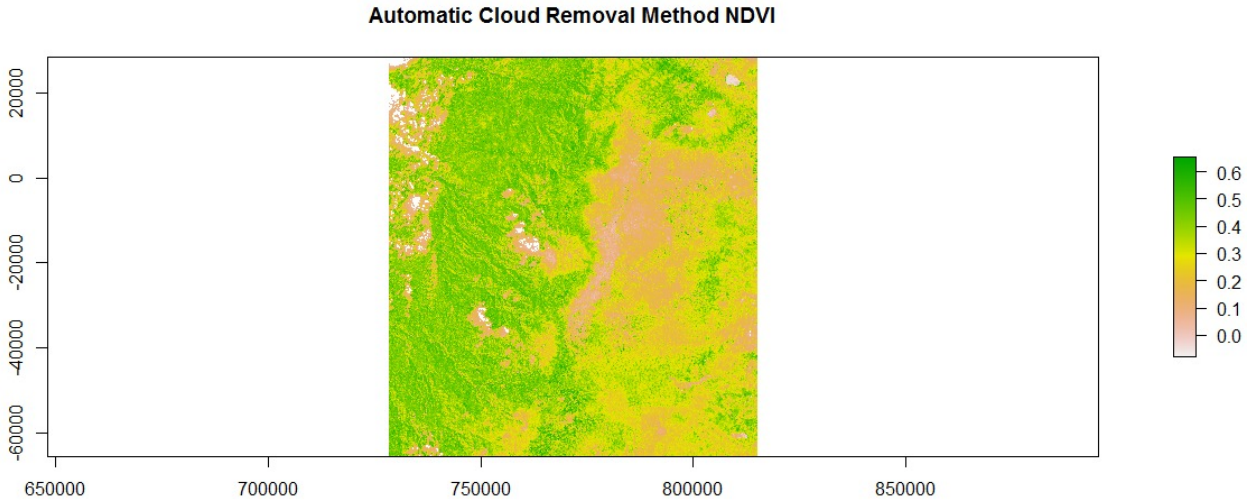


Figure 5. NDVI computation considering the Landsat 8 image after the application of ACRM algorithm for the study area.

3.3 Validating Data

In order to validate the ACRM algorithm, a MOD13Q1 product (NDVI 16-Day L3 Global 250 m version 6) was used as a reference data, resampled to a spatial resolution of 30 m (Figure 6), considering a similar period of the Landsat data used.

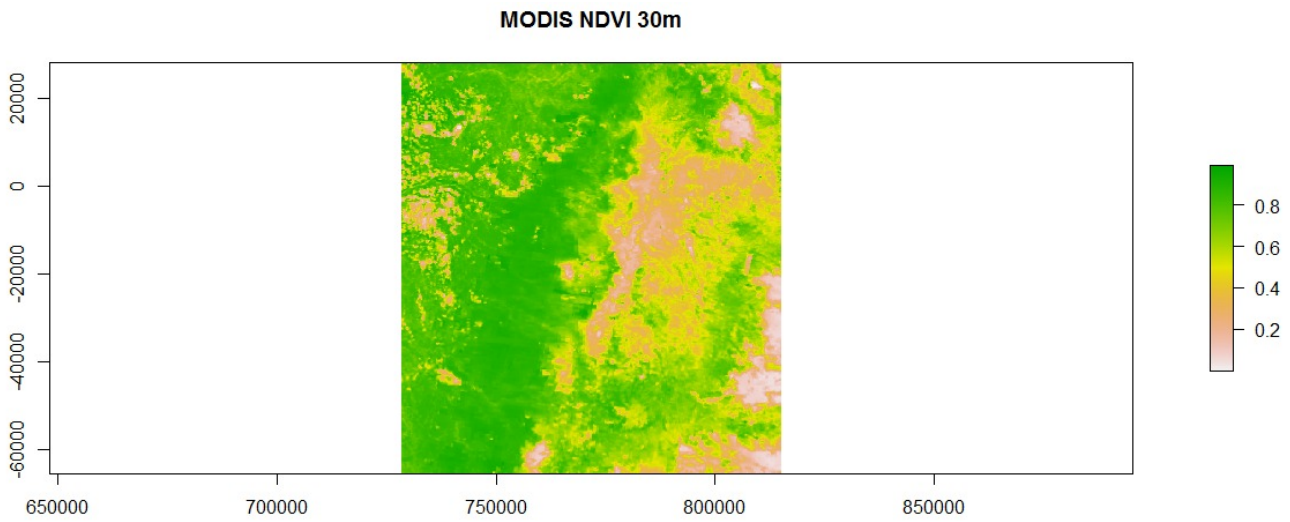


Figure 6. NDVI obtained from MOD13Q1 data for the study area (2014/07/28).

The idea to use MOD13Q1 is to compare the NDVI values with those values calculated with Landsat 8 original image (Figure 4) and Landsat 8 after applied ACRM algorithm (Figure 7). The validation was done over a small area that can be recognized (Quito Airport image of 2014/07/26) in an image with only a part of thin clouds to check the preliminary results considering the application of ACRM algorithm.

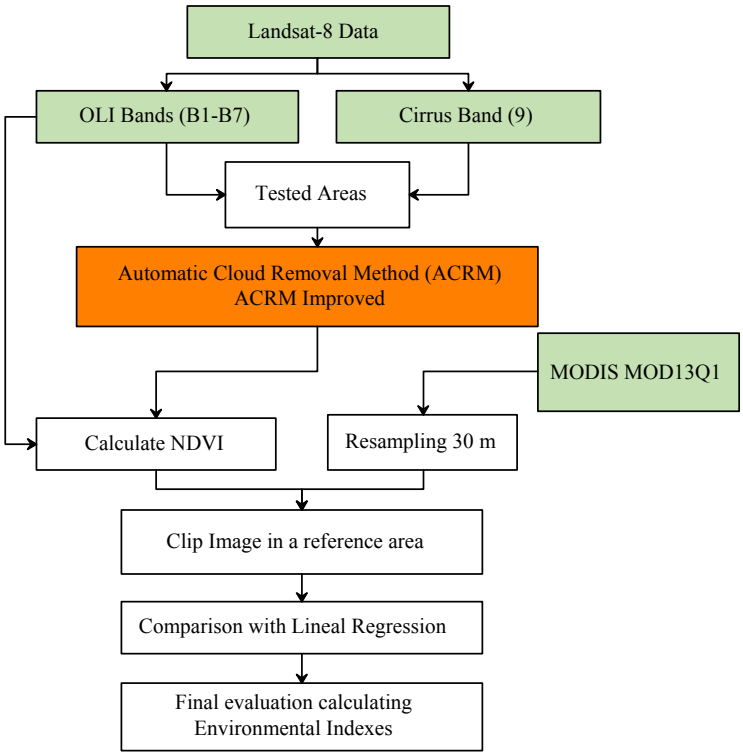


Figure 7. Workflow to evaluate ACRM (and Improved) algorithm in high elevation areas.

The ACRM algorithm tries to find the best slope fittest to the study area. Consequently, ACRM was validated in comparison with the product MODIS MOD13Q1 as a reference image because it had evaluated vegetation phenology in some studies[23,24] and it has been corrected[25] in comparison with NDVI calculated with original Landsat 8 imagery[26] and ACRM algorithm; moreover, MODIS product was resampled to 30 meters in spatial resolution to make the comparison[27] and validation. NDVI is used to validate because it is one of the most commonly used remote sensing indices[19,20] and can be considered an environmental index because it has relation with land surface dynamics[22].

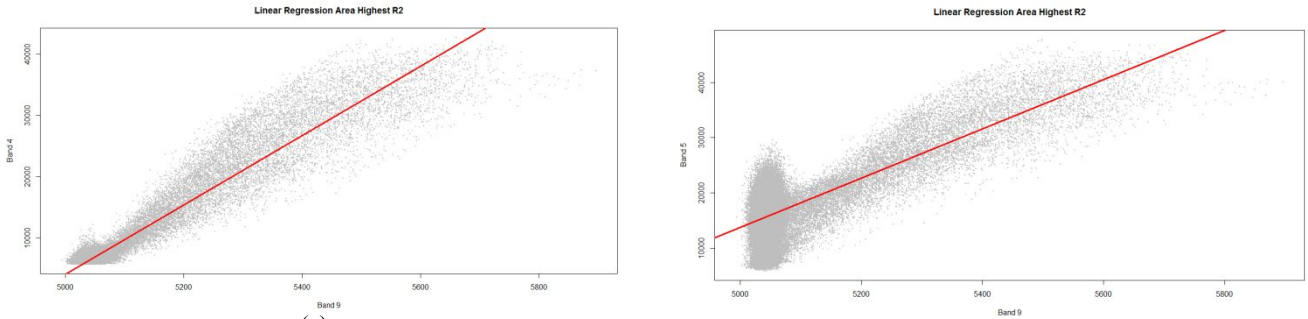
4. RESULTS AND DISCUSSION

4.1 Applying ACRM algorithm

Applying the ACRM algorithm over the 4 images in Quito the results shows (Table 1) a R^2 close to 1 (Figure 8), appears that algorithm works properly in this kind of regions. The special situation can be observed in the case of slope, where the values are closer to 0, considering that if the algorithm applies a value close to 0, this can be have a little correction and in comparison with original slopes obtained in Sidney, here the slopes are lower. In most of the cases, the slope between OLI Bands and Cirrus bands are close to 0. However, when we check visually the result, the algorithm does not work properly on the removal of thin cloud region, as can be observed in Figure 9.

Table 1. Coefficient of Determination (R^2) and Slope (α) obtained applying the ACRM algorithm in Quito.

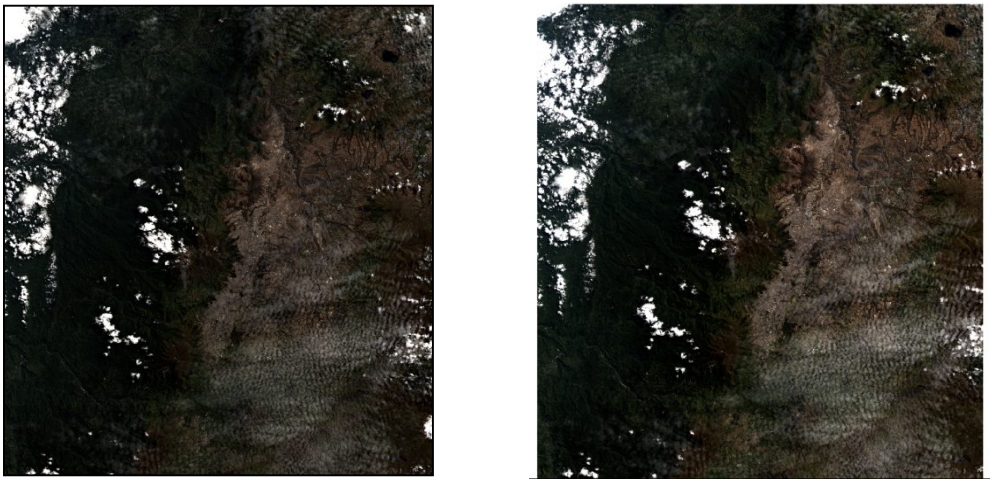
AREA	**QUITO (PATH:10 ROW:60)**		**QUITO (PATH:10 ROW:60)**		**QUITO (PATH:10 ROW:60)**		**QUITO (PATH:10 ROW:60)**	
DATE	**2013/07/07**		**2014/07/26**		**2015/08/30**		**2016/10/19**	
BAND	**R^2**	**Slope (α)**	**R^2**	**Slope (α)**	**R^2**	**Slope (α)**	**R^2**	**Slope (α)**
B2	0.96	0.05	0.93	0.02	0.97	0.03	0.95	0.03
B3	0.96	0.06	0.93	0.02	0.97	0.03	0.95	0.03
B4	0.95	0.05	0.93	0.02	0.97	0.02	0.96	0.03
B5	0.85	0.03	0.85	0.01	0.95	0.02	0.83	0.08
B6	0.90	0.06	0.89	0.17	0.92	0.02	0.91	0.04
B7	0.89	0.06	0.88	0.02	0.89	0.03	0.93	0.04



(a) (b)

Figure 8. Scatterplots considering the highest R^2 in the image: a) Linear Regression B9 vs. B4; b) Linear Regression B9 vs. B5.

Considering the lower slope values in Quito (Table 1), the ACRM algorithm was computed consider two differentes regions (Table 2). Different results in the evaluation were founded: Pedernales dind´t have a R^2 value close to 1; and Sidney (consider the same image used in the original algorithm) obtained a higher R^2, with a value close to 1 and a higher slope value when compared with all Quito images. Visually, it also can be checked that in Sidney image the ACRM algorithm works properlly, (Figure 10), but in Pedernales image the same is not true (Figure 11).

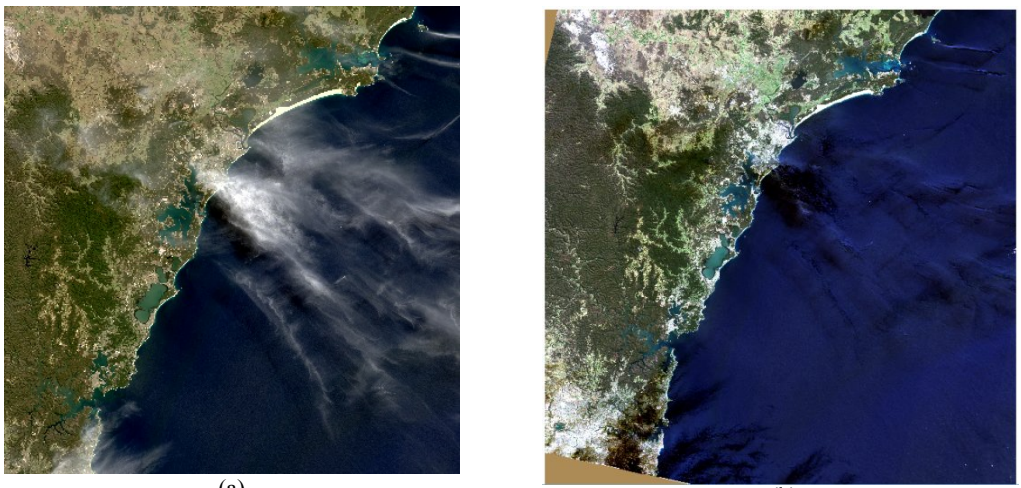


(a) (b)

Figure 9. Evaluation of ACRM algorithm in the Quito: a) Before apply ACRM; b) After apply ACRM.

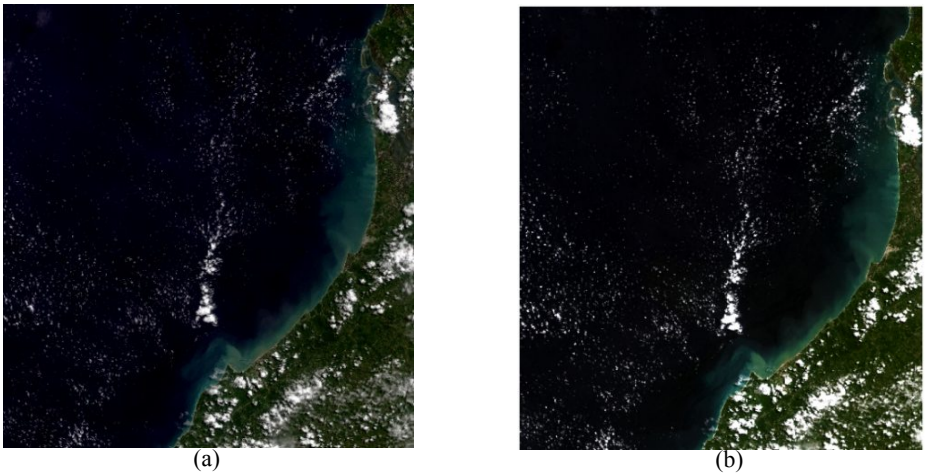
Table 2. Coefficient of Determination (R^2) and Slope (α) obtained applying ACRM in other places.

AREA	PEDERNALES (PATH:11 ROW:60)		SIDNEY (PATH:89 ROW:83)	
DATE	2016/05/13		2013/10/04	
BAND	R^2	Slope (α)	R^2	Slope (α)
B2	0.67	0.69	0.97	1.70
B3	0.68	0.68	0.99	1.63
B4	0.67	0.62	0.98	1.68
B5	0.67	0.52	0.98	1.74
B6	0.63	0.44	0.99	1.11
B7	0.53	0.58	0.98	1.02



(a) (b)

Figure 10. Evaluation of ACRM algorithm in the original Sidney image: a) Before apply ACRM; b) After apply ACRM.



(a) (b)

Figure 11. Evaluation of ACRM algorithm in Pedernales: a) Before apply ACRM; b) After apply ACRM.

4.2 Validating ACRM to calculate Environmental Indices

In order to validate the application of ACRM algorithm over Quito, we compared the NDVI from a reference image (MODIS MOD13Q1), the original Landsat image with the NDVI computation and the same Landsat image, but considering the application of the ACRM algorithm. With the application of ACMR algorithm the results can be satisfactory in order to recover pixel data, considering that other algorithms use masks and can lose the pixel value under thin clouds. The validation was realized in a small area in Quito where is located the airport and can be detected visually with remote sensing images. The R^2 between MODIS and the original Landsat 8 was 0.435, while the R^2 between MODIS and Landsat 8 consider the ACRM algorithm was 0.436. Therefore, this result shows an insignificant difference of R^2 in the retrieving of the environmental indices (NDVI).

4.3 Improving ACRM algorithm

Recognizing the contribution in the ACRM algorithm of the slope value, different slope values were tested in order to choose the best fitted in the region, considering only few changes considering the original algorithm values (1.988) obtained[13]. In this final consideration the Equation (5) shows how was applied the algorithm considering the slope change in order to improve and to know if this new conditions in slope value affects directly the quality of the results obtained.

$$x_i^f(u,v) = DN(u,v) - 1.988\left[c(u,v) - \min\{c(u,v)\}\right] \tag{5}$$

Visually, the result shows a substantial improvement in the removal on thin clouds (Figure 12).

(a) (b) (c)

Figure 12. a) Original Image; b) After apply ACRM algorithm; c) After ACRM algorithm Improved.

Also, the same method described in section 4.2 was considered here in order to validate the computation on NDVI in the new image after applied the ACRM algorithm (NDVI MODIS). The R^2 value is higher, with a value around 0.504 (Figure 13).

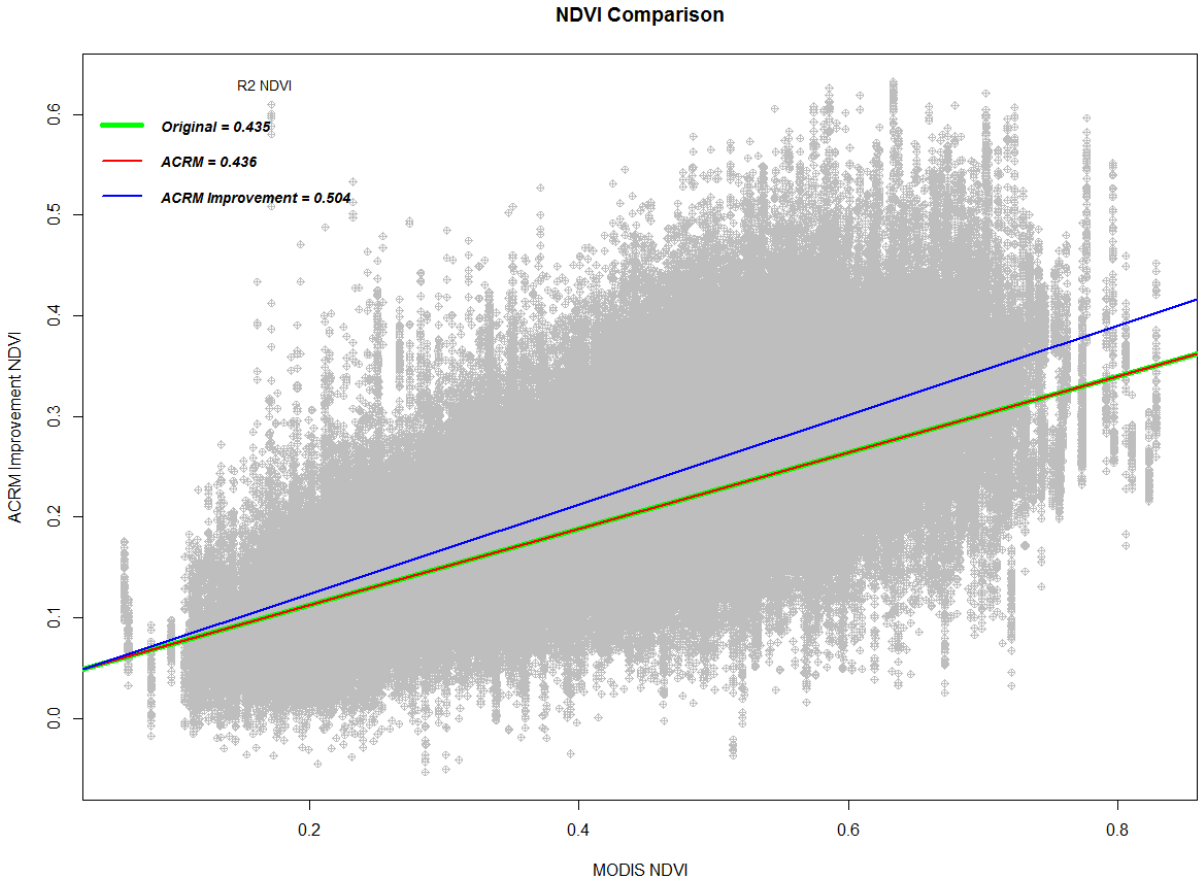


Figure 13. Lineal regression between MODIS NDVI and NDVI computed in Landsat 8 image, ACRM and ACRM Improved.

Improve the slope value in the ACMR algorithm allows to obtain better results and removal most of the thin clouds presented in the Landsat images. Also, was tested other images with the same slope values, but the results were the identical to those obtained in the original application of ACRM algorithm[13].

Accordingly, a simple modification in the slope of this algorithm improves in the data set to have a little bit better result in the final calculation of environmental indexes, having the challenge to try to find a way to improve the slope calculation in future studies.

5. CONCLUSIONS

The evaluation of ACRM algorithm in high elevation regions revealed that the ACRM algorithm cannot remove thin clouds in this kind of areas like Quito, Ecuador, considering that some factors like altitude and meteorological conditions can determine clouds presence during all the year. Accordingly, the application of the algorithm tries to remove thin clouds with the idea to don't lose the data under clouds to consider sensors like Landsat 8 to obtain environmental indices as NDVI. Landsat was considered because it has a higher spatial resolution when compared to other sensors, but a lower temporal resolution (16 days) trying to take advantage of the data obtained in each visit of this sensor. The results show a performance in the application when is calculated NDVI in comparison with a reference data like MODIS NDVI, obtaining a R^2 nearer to 1, but a low slope (nearer to 0). This final situation was to improve the algorithm finding the fittest slope to apply in this region, testing some slope values, considering that environmental indices can approach a reference. The point of view in future work is the question of what is the best way to obtain a fittest slope to improve this

algorithm and what is the fittest slope to each region around the world, considering that Cirrus band has a lot of possibilities to explain where the thin clouds are present.

ACKNOWLEDGEMENTS

This article was supported by Universidad Politécnica Salesiana Ecuador (UPS) and the scholarship to PhD programs offered by the UPS, by FEDER through the operation POCI-01-0145-FEDER-007690 funded by the Programa Operacional Competitividade e Internacionalização – COMPETE2020 and by National Funds through FCT – Fundação para a Ciência e a Tecnologia within ICT, R&D Unit (reference UID/GEO/04683/2013).

REFERENCES

[1] Rees, W. G., [Physical Principles of Remote Sensing], Cambridge University Press, (2012).
[2] Asner, G. P., "Cloud cover in Landsat observations of the Brazilian Amazon," Int. J. Remote Sens. **22**(18), 3855–3862 (2001).
[3] Fernández, G., Obermeier, W., Gerique, A., Sandoval, M., Lehnert, L., Thies, B. and Bendix, J., "Land Cover Change in the Andes of Southern Ecuador—Patterns and Drivers," Remote Sens. **7**(3), 2509–2542 (2015).
[4] Herring, J. W. and D., "Measuring Vegetation (NDVI & EVI)," 30 August 2000, <https://earthobservatory.nasa.gov/Features/MeasuringVegetation/measuring_vegetation_1.php> (12 July 2017).
[5] Department of the Interior U.S. Geological Survey., "Landsat 8 (L8) Data Users Handbook" (2016).
[6] USGS., "User Guide Landsat 8 Operational Land Imager (Oli)," 1–16 (2013).
[7] Cheng, Q., Shen, H., Zhang, L., Yuan, Q. and Zeng, C., "Cloud removal for remotely sensed images by similar pixel replacement guided with a spatio-temporal MRF model," ISPRS J. Photogramm. Remote Sens. **92**, 54–68 (2014).
[8] Lin, C.-H., Lai, K.-H., Chen, Z.-B. and Chen, J.-Y., "Patch-Based Information Reconstruction of Cloud-Contaminated Multitemporal Images," IEEE Trans. Geosci. Remote Sens. **52**(1), 163–174 (2014).
[9] Lv, H., Wang, Y. and Shen, Y., "An empirical and radiative transfer model based algorithm to remove thin clouds in visible bands," Remote Sens. Environ. **179**, 183–195 (2016).
[10] Wu, M., Wu, C., Huang, W., Niu, Z., Wang, C., Li, W. and Hao, P., "An improved high spatial and temporal data fusion approach for combining Landsat and MODIS data to generate daily synthetic Landsat imagery," Inf. Fusion **31**, 14–25 (2016).
[11] Shen, Y., Wang, Y., Lv, H.. and Qian, J., "Removal of thin clouds in Landsat-8 OLI data with independent component analysis, " Remote Sensing **7**(9), 11481–11500 (2015).
[12] Shen, Y., Wang, Y., Lv, H. and Li, H., "Removal of Thin Clouds Using Cirrus and QA Bands of Landsat-8," Photogramm. Eng. Remote Sens. **81**(9), 721–731 (2015).
[13] Xu, M., Jia, X. and Pickering, M., "Automatic cloud removal for Landsat 8 OLI images using cirrus band," Int. Geosci. Remote Sens. Symp.(September 2016), 2511–2514 (2014).
[14] Instituto Nacional de Meteorología e Hidrología., "Boletín Climatológico Anual 2015" (2016).
[15] R Core Team., "R: A Language and Environment for Statistical Computing" (2016).
[16] Hijmans, R. J., "raster: Geographic Data Analysis and Modeling" (2016).
[17] Bivand, R., Keitt, T. and Rowlingson, B., "rgdal: Bindings for the Geospatial Data Abstraction Library" (2016).
[18] Greenberg, J. A. and Mattiuzzi, M., "gdalUtils: Wrappers for the Geospatial Data Abstraction Library (GDAL) Utilities" (2015).
[19] Roy, D. P., Kovalskyy, V., Zhang, H. K., Vermote, E. F., Yan, L., Kumar, S. S. and Egorov, A., "Characterization of Landsat-7 to Landsat-8 reflective wavelength and normalized difference vegetation index continuity," Remote Sens. Environ. **185**, 57–70 (2016).
[20] Mishra, N. B. and Mainali, K. P., "Greening and browning of the Himalaya: Spatial patterns and the role of climatic change and human drivers," Sci. Total Environ. **587**–**588**, 326–339 (2017).

[21] Tucker, C. J., "Red and photographic infrared linear combinations for monitoring vegetation," Remote Sens. Environ. **8**(2), 127–150 (1979).
[22] Kuenzer, C., Dech, S. and Wagner, W., "Remote Sensing Time Series Revealing Land Surface Dynamics: Status Quo and the Pathway Ahead," Springer, Cham, 1–24 (2015).
[23] Sesnie, S. E., Dickson, B. G., Rosenstock, S. S. and Rundall, J. M., "A comparison of Landsat TM and MODIS vegetation indices for estimating forage phenology in desert bighorn sheep (*Ovis canadensis nelsoni*) habitat in the Sonoran Desert, USA," Int. J. Remote Sens. **33**(1), 276–286 (2012).
[24] Liu, Y., Hill, M. J., Zhang, X., Wang, Z., Richardson, A. D., Hufkens, K., Filippa, G., Baldocchi, D. D., Ma, S., Verfaillie, J. and Schaaf, C. B., "Using data from Landsat, MODIS, VIIRS and PhenoCams to monitor the phenology of California oak/grass savanna and open grassland across spatial scales," Agric. For. Meteorol. **237**–**238**, 311–325 (2017).
[25] Solano, R., Didan, K., Jacobson, A. and Huete, A., "MODIS Vegetation Index User ' s Guide (MOD13 Series)" (2010).
[26] Ke, Y., Im, J., Lee, J., Gong, H. and Ryu, Y., "Characteristics of Landsat 8 OLI-derived NDVI by comparison with multiple satellite sensors and in-situ observations," Remote Sens. Environ. **164**, 298–313 (2015).
[27] Zhang, H. K. and Roy, D. P., "Using the 500 m MODIS land cover product to derive a consistent continental scale 30 m Landsat land cover classification," Remote Sens. Environ. **197**, 15–34 (2017).

Comparison of satellite remote sensing data in the retrieve of PM10 air pollutant over Quito, Ecuador

Cesar I. Alvarez-Mendoza, Ana Teodoro, Nelly Torres, Valeria Vivanco, Lenin Ramirez-Cando

Cesar I. Alvarez-Mendoza, Ana Teodoro, Nelly Torres, Valeria Vivanco, Lenin Ramirez-Cando, "Comparison of satellite remote sensing data in the retrieve of PM10 air pollutant over Quito, Ecuador," Proc. SPIE 10793, Remote Sensing Technologies and Applications in Urban Environments III, 107930I (9 October 2018); doi: 10.1117/12.2325324

SPIE.

Event: SPIE Remote Sensing, 2018, Berlin, Germany

Comparison of satellite remote sensing data in the retrieve of PM10 air pollutant over Quito, Ecuador

Cesar I. Alvarez-Mendoza[*a,b], Ana Teodoro[a,c], Nelly Torres[b], Valeria Vivanco[b], Lenin Ramirez-Cando[b]

[a]Dep.of Geosciences, Environment and Land Planning, Faculty of Sciences, University of Porto, 4169-007 Porto, Portugal; [b]Universidad Politécnica Salesiana, Grupo de Investigación Ambiental en el Desarrollo Sustentable GIADES, Carrera de Ingeniería Ambiental, Quito, Ecuador; [c]Earth Sciences Institute (ICT), Pole of the FCUP, University of Porto, Portugal.

ABSTRACT

Most of the large cities have an air quality network to measure air pollution including PM10. However, air quality monitoring network has a high cost and it is spatially limited. Quito, capital of Ecuador, is a city with an automatic air quality network (REMMAQ) composed by 9 stations. The REMMAQ works since 2002, measuring PM10 only in 4 regular stations located at different points along the city. This scarce quantity of PM10 measures led us to propose a new strategy to obtain PM10 data in all the city. Several studies have already considered the retrieving of PM10 from remote sensing data in cities with an air quality network. In order to find an optimal model to retrieve PM10 in Quito, this study compare the use of 3 different satellite sensors (Landsat-7 ETM+, Landsat-8 OLI and TERRA/MODIS) between 2013 to 2017. Additional to remote sensing data, we also use field data considering the REMMAQ. In each sensor, we used different variables and environmental indexes to model the best fit equation to quantify PM10 in all the city, finding the significant variables for each type of data and year. The variables considered were the Normalized Difference Vegetation Index (NDVI), Land Surface Temperature (LST), Soil-adjusted Vegetation Index (SAVI), Normalized Difference Water Index (NDWI), Normalized Stability Index (NSI), surface reflectance Blue Band (B1), surface reflectance Green Band (B2) and surface reflectance Red Band (B3). These variables were considered because most of them are used in different studies combined with meteorological data. All the procedures were implemented in R Studio. The empirical models using remote sensing data/derived products and air quality data can help in retrieving air pollutants in large cities. This work is a valuable contribution for the study of the spatialization of PM10 in order to find new alternatives in the use of remote sensing data to support government decisions.

Keywords: PM10, Landsat, MODIS, Air Quality, Quito

1. INTRODUCTION

One of the changes that Earth had suffered on its dynamic is the air quality, where human activities as car traffic, industries and other activities generate air pollution[1]. Air pollution includes gaseous and particulate contaminants considering the last as a problem in the respiratory human health[2]. The World Health Organization (WHO) claims that most people around the world are breathing air polluted, specifically particulate contaminants[3]. Within particulate contaminants, one of the most common is particulate matter of less than 10 microns (PM10). PM10 is a pollutant that can be measure by air quality station in the cities[4]. Most of the largest capital cities have an air quality monitoring network (AQMN) to measure air pollution including PM10. However, acquiring an air quality station can cost a lot of money and it results limiting to municipal governments[5]. Quito, the capital of Ecuador, is a city with an AQMN. Its name is Red Metropolitana de Monitoreo Atmosférico de Quito (REMMAQ) who is composed by 9 stations[6]. The REMMAQ works since 2002, where PM10 is measured only in some years by nine automatic stations. The low quantity of PM10 measures requires a different strategy to obtain data with more accuracy in all the city, especially in the urban part[7]. Several studies consider the retrieving of PM10 from remote sensing data in cities with an air quality network[8–11]. In the case of Quito, a previous study shows good results with Landsat-7 images[8]. Other studies around the world used Landsat-8 and Moderate Resolution Imaging Spectroradiometer (MODIS) data using empirical models [9–11]. In the case of empirical models to retrieve air

[*]calvarezm@ups.edu.ec; phone +593 9 84647745

Remote Sensing Technologies and Applications in Urban Environments III, edited by Thilo Erbertseder, Nektarios Chrysoulakis, Ying Zhang, Proc. of SPIE Vol. 10793, 107930I · © 2018 SPIE
CCC code: 0277-786X/18/$18 · doi: 10.1117/12.2325324

quality data from remote sensing data, some of the environmental indexes are computed in order to retrieve PM10. The multiple linear regression (MLR) considering the visible bands is one of the most frequently method applied to estimate air quality with remote sensing data[12]. Nevertheless, due to this methodology cannot treat intercorrelated variables and missing data[13], a Partial Least Square (PLS) methodology can be used in order to analyze the collinearity between spatial data[14].

In order to find an optimal and accuracy model to retrieve PM10 in Quito, this study compare three different remote sensing data (Landsat 7 ETM+, Landsat 8 OLI and TERRA/MODIS) between 2013 to 2017. Where, additional to remote sensing data, the use of field measurements is important considering the REMMAQ. The model proposed is built considering a PLS Regression[15].

2. STUDY AREA AND DATA

2.1 Study area

Quito is the capital of Ecuador. It is a city with some air pollutions problems, as many cities around the world. One of this air pollution problems is the car traffic[16]. Quito is located above equatorial line, in the middle of Andean regions in South America[17], where, the meteorological conditions can strongly influence the air quality. For this project, the study area is centered in the urban region in Quito, where the REMMAQ stations are presented (Figure 1).

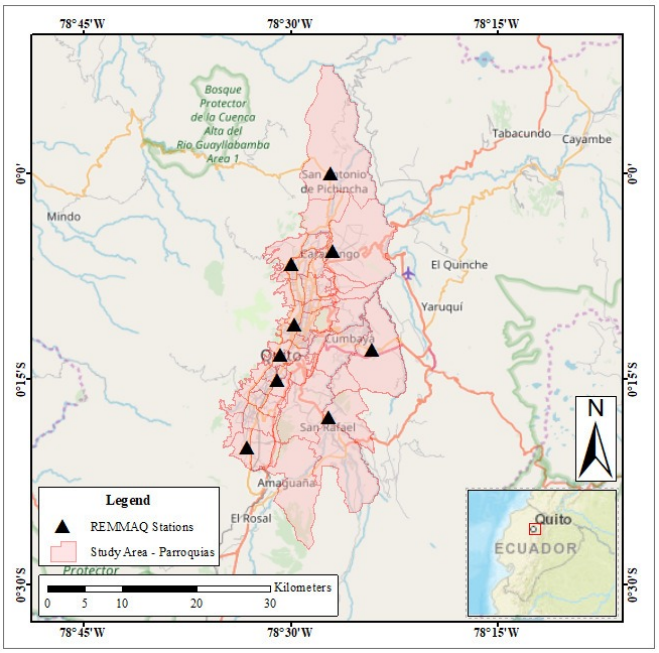


Figure 1. Study area location (Quito, Ecuador). The black points are the REMMAQ Stations. The red polygons are the districts or "parroquias". Most of the stations are located in the urban area.

2.2 Data collection

About the remote sensing data, three satellite sensor data were used (Landsat-7 ETM+, Landsat-8 OLI and TERRA/MODIS) between 2013 to 2017 (Figure 2). The images were obtained from Earth Explorer website[18]. These images are open access and freely download. In the case of Landsat data, it was chosen considering the spatial resolution[19,20] (30 m) and also because most of the similar studies around the world use this type of data. On the other hand, MODIS was selected considering the temporal resolution and the availability of ready products to use[21]. Only images with less than 10% clouds in the study area were considered. A total of approximately 30 images for each sensor was considered. In all the data the surface reflectance ready products were used.

In order to compute a mathematical empirical model, the air quality data was obtained by REEMAQ stations considering the same period of remote sensing data (2013 – 2017). The data were downloaded from Secretaria del Ambiente Website[22]. Unfortunately, just three stations have PM10 measures during all the studied period. These data are public.

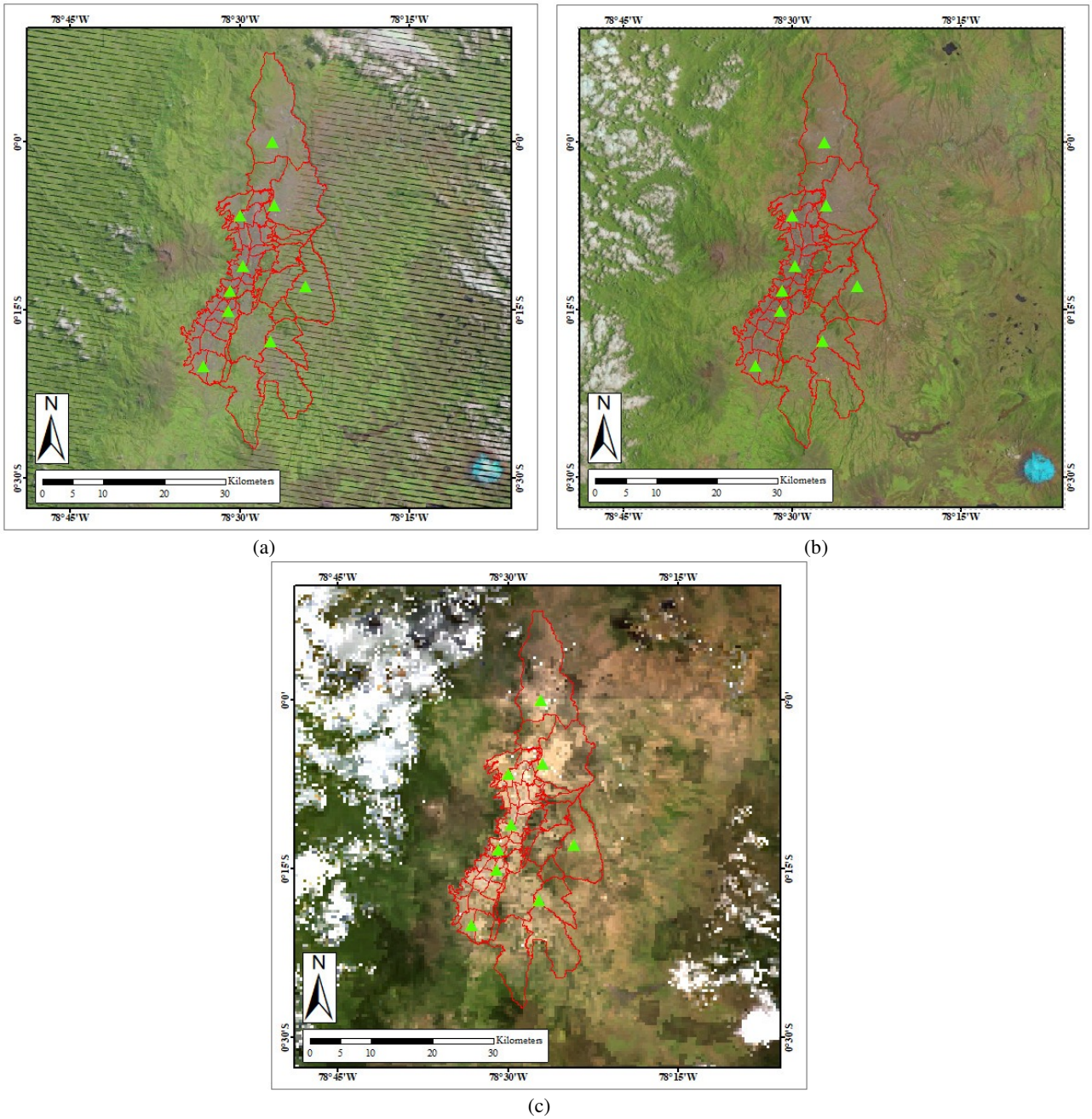


Figure 2. Example of images used to compute the PM10 model in the study area (Red color polygons): (a) Landsat-7 ETM+; (b) Landsat-8 OLI; (c) MODIS

3. METHODOLOGY

In order to generate a model to retrieve PM10, remote sensing data and field air data were considered. Firstly, database containing the most adequate satellite bands (visible, NIR, SWIR bands), some environmental indexes and PM10 measurements was built. With this database, PLS regression was applied in order to compute the empirical models to each sensor. Finally, the models were applied in the images and the PM10 was retrieving and mapped (Figure 3).

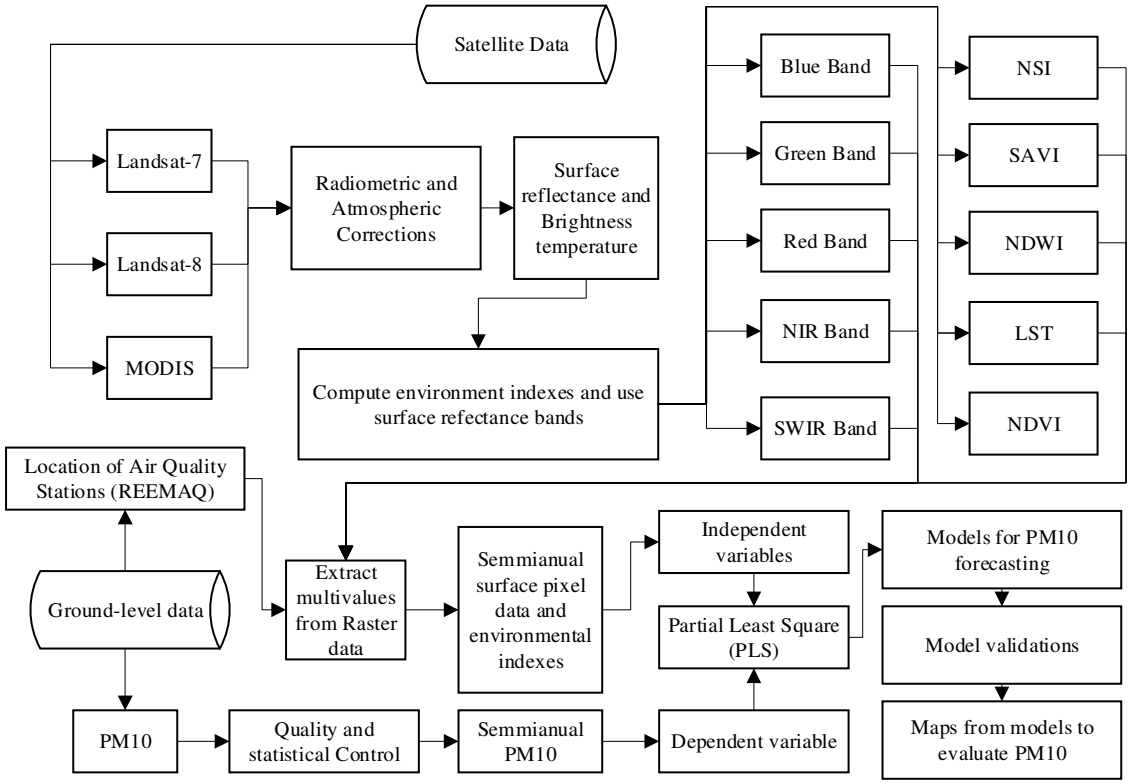


Figure 3. Methodology workflow.

The surface reflectance data from Blue, Green, Red, Near Infrared (NIR) and Short-wave infrared (SWIR) bands were extracted from each image. Besides, some of the most common environmental indexes as Normalized Difference Vegetation Index (NDVI), Normalized Difference Soil Index (NDSI), Soil-Adjusted Vegetation Index (SAVI), Normalized Difference Water Index (NDWI) and Land Surface Temperature (LST) were used in this study as independent variables. A point shapefile with the location of REMMAQ stations was used to extract the raster values. The ArcGIS 10.5 with Extract Multi values to points tool was used to extract the data in each station[23].

3.1 Surface reflectance data

The Blue, Green, Red, NIR and SWIR Bands were extracted from each sensor. These variables are used because most of the studies in similar regions prove to be a relation between visible and infrared data with PM10[8,24–26]. Moreover, the aerosol optical thickness (AOT) or aerosol optical depth (AOD) is a measurement of the aerosols, where PM10 can be contained[27]. The AOT shows how the atmosphere reflects and absorbs visible and infrared light[28].

Landsat program has acquired land surface data since 1972. Nowadays, Landsat-7 and Landsat-8 are operational, obtaining visible, infrared and thermal data in a middle spatial and temporal resolution[29]. In order to use surface data and computing environmental indexes, the Landsat surface data Level-2 were download to each image. The advantage of Level-2 is to have data ready to use with all the corrections applied (Geometric, radiometric and atmospheric corrections)[30,31].

MODIS sensor is presented in Terra and Aqua satellites. Its uses are in the land surface observation, aerosol detecting, etc. An advantage of this satellite is to have ready products (Geometric, radiometric and atmospheric corrections) as surface bands (MOD09)[32,33] and some environmental indexes. MODIS has 36 bands and a high temporal resolution.

3.2 Normalized Difference Vegetation Index (NDVI)

The NDVI is an environment index that allowed to obtain information about the greenest vegetation, using Red and NIR bands[34]. In order to consider the influence of vegetation over PM10 in the urban areas[35], the NDVI was computed to Landsat-7 and Landsat-8 data, as shown in Equation (1).

$$NDVI = \frac{NIR-RED}{NIR+RED} \quad (1)$$

On the other hand, MOD13Q1[36] product was used to get NDVI data from MODIS. This product has a 250 m of spatial resolution. The pixel data were multiplied by 0.0001[36].

3.3 Normalized Difference Soil Index (NDSI)

The NDSI index was computed considering surface reflectance data. It identifies zones where built areas are presented[37]. NDSI is computed by Equation (2).

$$NDSI = \frac{SWIR-NIR}{SWIR+NIR} \quad (2)$$

3.4 Soil-Adjusted Vegetation Index (SAVI)

The SAVI is an improvement of NDVI, where a soil correction factor is introduce to prevent the reduction of difference in Red and NIR of the canopy by background soil[38], as shown in Equation (3).

$$SAVI = (1 + L)\frac{NIR-RED}{NIR+RED+L} \quad (3)$$

Where, L value is 0.5, considering that the change in soil brightness is minimal.

3.5 Normalized Difference Water Index (NDWI)

The NDWI maximizes the reflectance of water by using Green and NIR bands (surface reflectance). The aim is to build a model with water consideration[39]. It is expressed as shown in Equation (4).

$$NDWI = \frac{GREEN-NIR}{GREEN+NIR} \quad (4)$$

3.6 Land Surface Temperature (LST)

In order to compare PM10 with meteorological data, the LST was obtained from thermal bands by Inversion of Planck's function[40] in order to become a variable in the model. The LST is in Kelvin degrees. Converting to Celsius degrees requires to subtract 273.15 value. LST computation is described by equation (5).

$$LST = \frac{BT}{\left(1+\left(\frac{\lambda * BT}{\rho}\right) ln\varepsilon\right)} \quad (5)$$

Where, BT is the brightness temperature obtained from Landsat Level 2 products. λ is the center wavelength (Landsat-7 = 11.45 μm, Landsat-8 = 10.8 μm)[41], ρ is the a constant obtained as Equation (6) and ε is the emissivity as Equation (7).

$$\rho = \frac{h*c}{s} \quad (6)$$

Where, h is the Planck's constant (6.626e-34 Js), c is the velocity of light (2.998e8 m/s) and s is the Boltzmann constant (1.38e-23 J/K).

Furthermore, the emissivity is a variable required to compute LST. It is defined as the efficiency with a surface emits heat as Thermal Infrared (TIR) radiation[42]. The algorithm showed by Equation 7 considered a semi-empirical method where the variations of NDVI in the vegetation ($NDVI_v$) and soil ($NDVI_s$) are important[43]. In order to choose the item in the Equation 7, the NDVI should be analyzed in the study area.

$$\varepsilon = \begin{cases} \varepsilon_s, \ NDVI < NDVI_s \\ \varepsilon_s + (\varepsilon_v - \varepsilon_s)P_V, \ NDVI_s \leq NDVI \leq NDVI_v \\ \varepsilon_v, \ NDVI > NDVI_v \end{cases} \quad (7)$$

Where, ε_s represents the emissivity for soil considering in this study a value of 0.973. ε_v is the emissivity for vegetation considering a value of 0.985[44] and P_v is the proportion of vegetation in the area, which is computed by Equation 8, using the NDVI.

$$P_V = \left(\frac{NDVI - NDVI_s}{NDVI_v - NDVI_s}\right)^2 \quad (8)$$

$NDVI_v$ and $NDVI_s$ were 0.2 and 0.5 values, respectively, used in Equation 8[43].

MOD11A2 is the product used to retrieve LST in MODIS sensor. Its pixel size is 1000 meters. The scale factor used to get LST was 0.02[45].

3.7 Air Quality Measurements

PM10 measurements were obtained from REEMAQ Stations. The REEMAQ works since 2002 with automatic stations. The stations measure some air pollutants as CO, SO2, NOx, O3, PM2.5 and PM10 in an hourly basis. One of the major challenges was to retrieve the PM10 data from stations because we found them only in some years and in ten stations (Table 1).

Table 1. PM10 Semmianual median by each REEMAQ Station founded between 2013 – 2017

REEMAQ Station	Years data	PM10 Semmianual median (µg/m^3)
Belisario	2013 – 2016	31.9
Carapungo	2013 – 2017	84.1
Cotocollao	2013 – 2014	33.8
El Camal	2013	60.9
Guamaní	2013 – 2017	40.1
Jipijapa	2014 – 2016	58.8
Los Chillos	2013 – 2016	27.3
San Antonio	2017	54.6
Tababela	2013 – 2016	35.8
Tumbaco	2013 - 2017	42.9

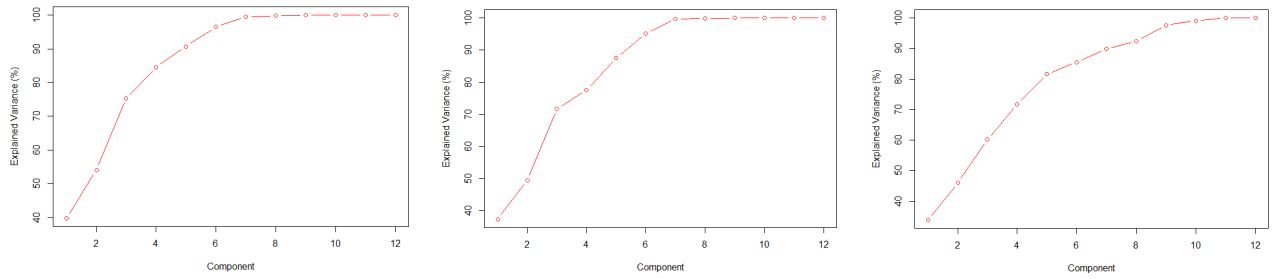
Due to the few quantities of PM10 field data, the semiannual median was used to build the variable database. In this database was considered remote sensing data, PM10 field measurements, and additional data as Season (Season 1 January to June or Season 2 July to December) and Year. Moreover, the PM10 semiannual medians data were obtained with hourly data measurements. The hourly data collected were between 10:00 to 11:00 (GMT-5) in each station according to the time when Landsat-8 acquires data.

A PLS regression was employed in order to predict the dependent value (PM10) from a set of predictors. This technique is used to handle a possible multicollinearity. Likewise, PLS regression can be used when standard regression methods fail, and we have multiple data collected on the same observations[15]. R studio was the software used to compute PLS regression, considering the package *pls* and *plsdepot*.

4. RESULTS AND DISCUSSION

The final semiannual tables generated to each sensor contains 29 observations for Landsat-7, 53 observations for Landsat-8 and 59 observations for MODIS. Applying the PLS technique in each semiannual data table, the first part was to analyze the number of components in the model. The aim of PLS is to explain the variance model with the less quantity of

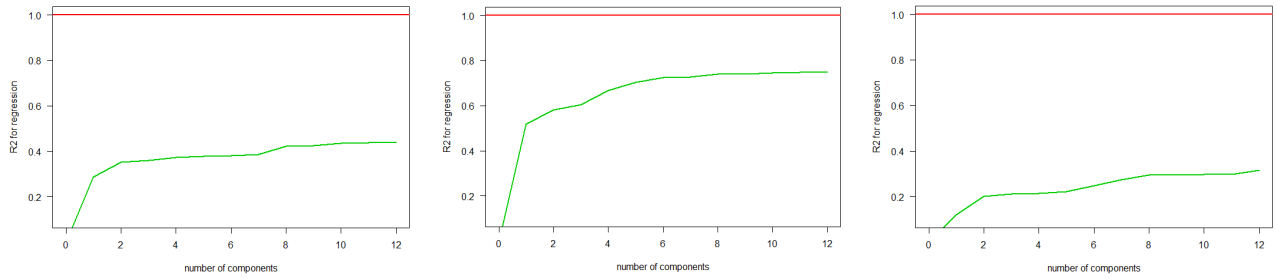
components (Figure 4). In the case of Landsat-7 and Landsat-8 was considered 10 components to explain the model variance. On the other hand, MODIS considered 12 components in the PLS model.



(a) (b) (c)

Figure 4. Explained Variance Vs. Number of components in PLS Regression (a) Landsat-7; (b) Landsat-8; (c) MODIS

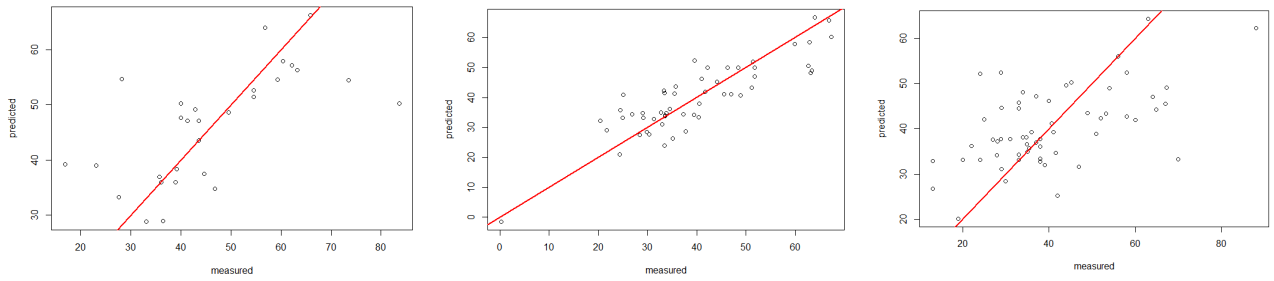
In order to choose the fittest model, the R^2 was analyzed, for each sensor (Figure 5). For Landsat-7 was founded a value of 0.41, for Landsat-8 a value of 0.72 and for MODIS a value of 0.28. Considering the R^2 values obtained to retrieve PM10, the fittest model is considering Landsat-8 (Figure 5).



(a) (b) (c)

Figure 5. Coefficient of Determination with components (R^2) (a) Landsat-7; (b) Landsat-8; (c) MODIS

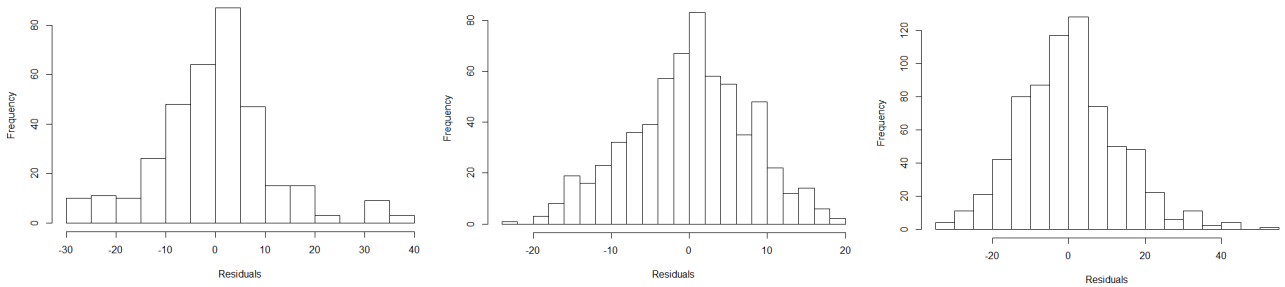
The model validation was done with a comparison between predicted vs. measured PM10 (Figure 6) and a histogram of residuals (Figure 7), where the model considered in the linear equation fittest is the Landsat-8 model. Additionally, in the Landsat-8 model the residual shows a trend of a normal distribution with residual values until 20 $\mu g/m^3$.



(a) (b) (c)

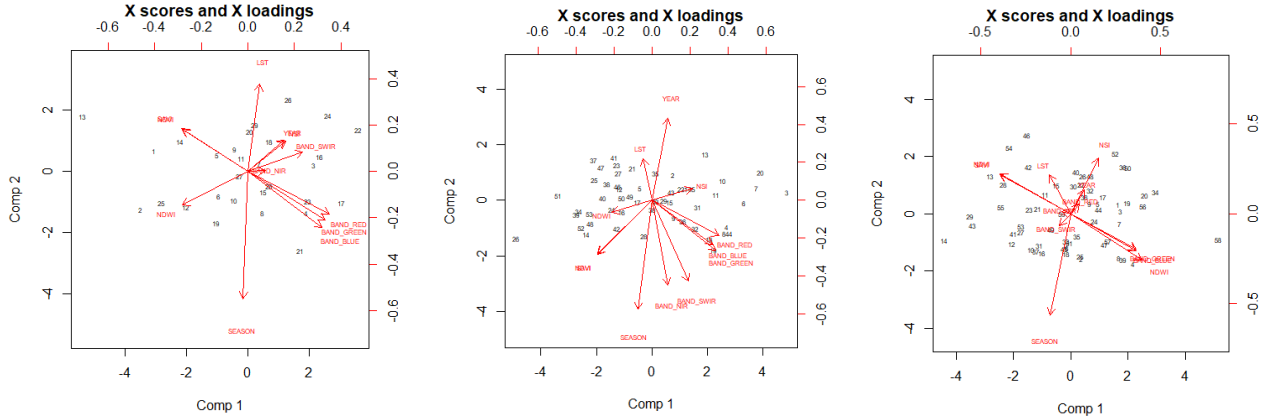
Figure 6. Predicted values Vs. Measured values (a) Landsat-7; (b) Landsat-8; (c) MODIS

In the models is evident the multicollinearity (vectors that follow the same direction and magnitude). Therefore, plotting individual factor scores, allows to identify the most relevant variables in projection of the information over the new latent variables. These new variables (all Comps), however, contains information for variance in dependent variable that will be used later to model its behavior taking the information from the components considering a regression equation (Figure 8).



(a) (b) (c)

Figure 7. Histogram of residuals (a) Landsat-7; (b) Landsat-8; (c) MODIS



(a) (b) (c)

Figure 8. Biplot of data used in PLS regression (a) Landsat-7; (b) Landsat-8; (c) MODIS

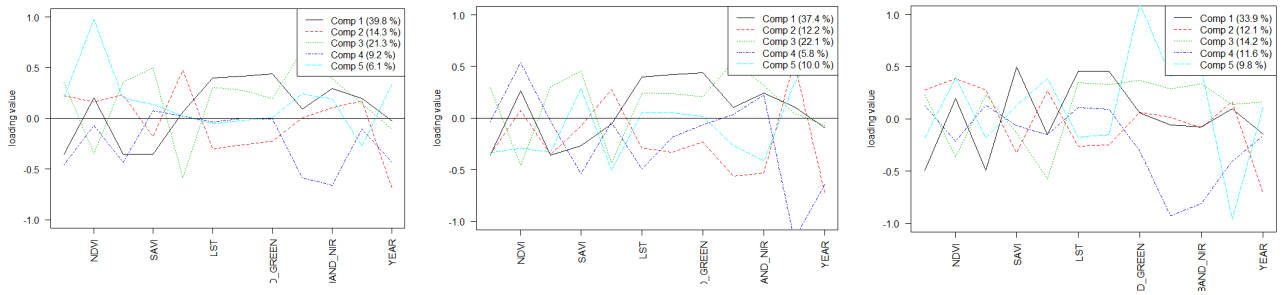
PLS regression, in fact, used correlated variables information to create new "variables" called components that are uncorrelated, increasing the reliability of the model. In the three sensors, we can see that the first component gives the most percentage of variance explication (Figure 9).

The Equation 9 shows the final model, where the remotes parameters were taken as independent variables.

$$PM10 = I + aNDVI - bNSI - cSAVI + dNDWI - eLST - fB - gG + hR + iN + jSW + kYEAR - lS \quad (9)$$

Where, I is the intercept, B is the blue band, G is the green band, R is the red band, N is the NIR band, SW is the SWIR band, s is the season, a, b…, l are the coefficients to each independent variable.

The values of intercept and coefficients were computed with PLS (Table 2) to each sensor. The fittest equation was obtained for Landsat-8 with the highest R^2 value (0.74).



(a) (b) (c)

Figure 9. Components loadings on variables (X matrix) (a) Landsat-7; (b) Landsat-8; (c) MODIS

Table 2. Model coefficients (considering Equation 9)

Coefficient	Landsat-7	Landsat-8	MODIS
Intercept (I)	-6385.47	-3327.619	-97.9388
NDVI - a	-137.7962	255.9988	8.4164
NSI - b	55.7566	-13.3961	174.1538
SAVI - c	31.0459	-236.1787	44.3324
NDWI - d	-62.9931	13.3961	57.7680
LST - e	0.6050	-1.6362	0.5463
BAND_BLUE - f	-0.0680	-0.0477	0.0936
BAND_GREEN - g	0.1531	-0.0408	-0.0746
BAND_RED - h	-0.0908	0.0735	0.0000089
BAND_NIR - i	0.0378	0.0082	0.0000005
BAND_SWIR - j	-0.0176	0.0048	0.0000004
YEAR - k	3.1807	1.6918	0.0632
SEASON - l	-6.9296	-3.8854	-8.6446
R^2	0.41	0.74	0.28

Applying the model presented in Equation 9 to the Landsat-8 images (Figure 10), the resulting raster shows the PM10 concentration in all the study area for the image date.

The limitation of the PM10 algorithm is directly related to the images quality (clouds), images availability and PM10 field measurements. In the case of Quito, most of the images have more than 20% of clouds and the REEMAQ stations are not constant during the study time. This was the main reason why we choose semiannual medians to input variables in the model. One of the main innovations of this work is the consideration of Landsat-8 images (30 images). It is the higher difference with other similar studies around the world where the R^2 obtained in these models is lower and uses less quantity of data[10,46]. In some studies, meteorological ground data[47] are also used to retrieve PM10 , but these variables are too difficult to obtain. Other studies uses only MODIS products to retrieve PM10[48,49], but this sensor has a low spatial resolution, which can be a limitation. In addition, in this work, the PLS regression was chosen to avoid a possible correlation between the independent variables.

In order to retrieve PM10 in Quito between 2013 – 2017, the Landsat-8 model can be used to obtain better results than other sensors as Landsat-7 or MODIS. The reasons can be: (1) the image quality in comparison with Landsat 7 SLC-off[50]; (2) the higher spatial resolution. This model can be applied to all the images between 2013 to 2017, generating new PM10 concentration maps that could be used for the governmental authorities to take decisions.

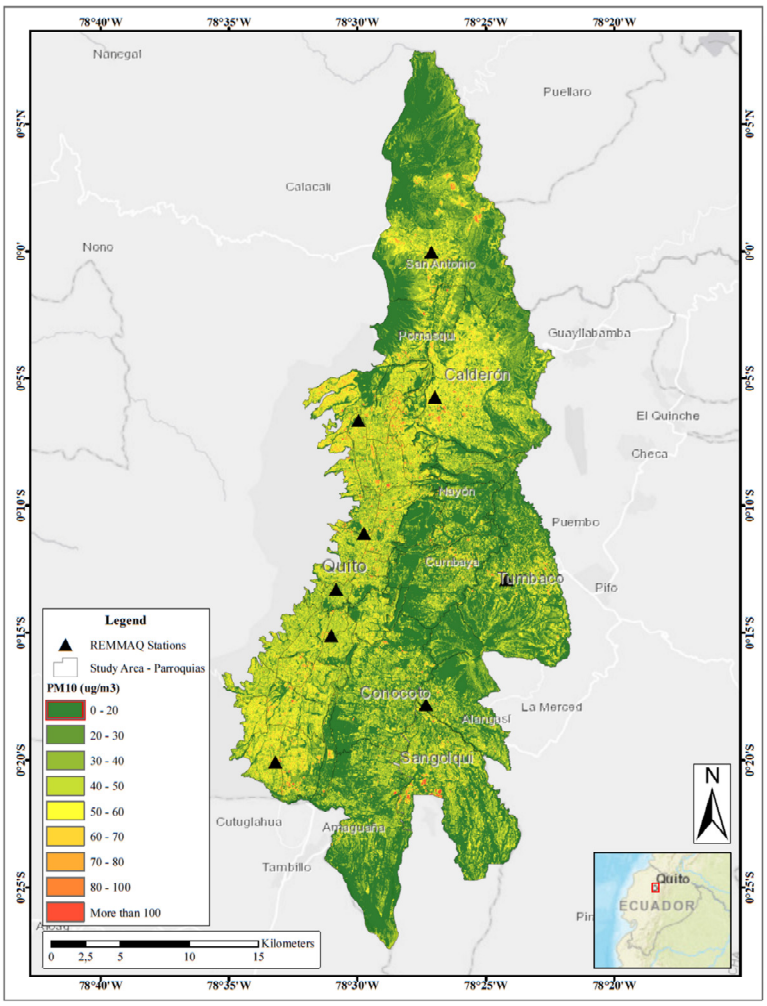


Figure 10. Landsat-8 PM10 retrieved (20/09/2018).

5. CONCLUSIONS

In this study, three satellite sensors were considered in order to retrieve PM10 from remote sensing data, in Quito, Ecuador. The evaluation showed than Landsat-8 images give the fittest model ($R^2 = 0.74$) in comparison with Landsat-7 ($R^2 = 0.41$) and MODIS ($R^2 = 0.28$). PLS regression was used to compute the models to retrieve PM10. This is a robust technique that discompose the original predictors values in components. Its results are useful when just a few PM10 field measurements/observations are available in different periods. The model was applied in all the Landsat-8 images between 2013 to 2017 available in the dataset, showing the behavior of PM10 during this period. Also, using the model proposed in this study is possible to find a possible relation with respiratory diseases cases in some places in Quito. As a future work, the work will use more regression techniques to improve the results.

REFERENCES

[1] AAAS Project 2061., "Communicating and learning about global climate change," 2007, <https://climate.nasa.gov/resources/education/pbs_modules/lesson1Engage/> (25 June 2018).

[2] Karagulian, F., Belis, C. A., Dora, C. F. C., Prüss-Ustün, A. M., Bonjour, S., Adair-Rohani, H. and Amann, M., "Contributions to cities' ambient particulate matter (PM): A systematic review of local source contributions at global level," Atmos. Environ. **120**, 475–483 (2015).

[3] WHO., "9 out of 10 people worldwide breathe polluted air, but more countries are taking action," 2018, <http://www.who.int/news-room/detail/02-05-2018-9-out-of-10-people-worldwide-breathe-polluted-air-but-more-countries-are-taking-action> (25 June 2018).
[4] APCD., "5-year Air Quality Monitoring Network Assessment," San Diego (2015).
[5] Kumar, P., Morawska, L., Martani, C., Biskos, G., Neophytou, M., Di Sabatino, S., Bell, M., Norford, L. and Britter, R., "The rise of low-cost sensing for managing air pollution in cities," Environ. Int. **75**, 199–205 (2015).
[6] Secretaría de Ambiente., "Informe de la calidad de aire-2016" (2017).
[7] Munir, S., Gabr, S., Habeebullah, T. M. and Janajrah, M. A., "Spatiotemporal analysis of fine particulate matter (PM2.5) in Saudi Arabia using remote sensing data," Egypt. J. Remote Sens. Sp. Sci. **19**(2), 195–205 (2016).
[8] Álvarez Mendoza, C. I. and Padilla Almeida, O., "Estimación de la contaminación del aire por PM10 en Quito a través de índices ambientales con imágenes LANDSAT ETM+.," Rev. Cart.(92), 135–147 (2016).
[9] Nguyen, N. H. and Tran, V. A., "Estimation of PM10 from AOT of satellite Landsat 8 image over Hanoi City," Int. Symp. Geoinf. Spat. Infrastruct. Dev. Earth Allied Sci. (2014).
[10] Luo, N., Wong, M. S., Zhao, W., Yan, X. and Xiao, F., "Improved aerosol retrieval algorithm using Landsat images and its application for PM10 monitoring over urban areas," Atmos. Res. **153**, 264–275 (2015).
[11] Zhai, L., Sang, H., Zhang, J. and An, F., "Estimating the spatial distribution of $PM_{2.5}$ concentration by integrating geographic data and field measurements," ISPRS - Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci. **XL-7/W4**(July), 209–213 (2015).
[12] Chen, Y., Han, W., Chen, S. and Tong, L., "Estimating ground-level PM2.5 concentration using Landsat 8 in Chengdu, China," Proc. SPIE **9259**(February), 925917–925931 (2014).
[13] Fragkaki, A. G., Tsantili-Kakoulidou, A., Angelis, Y. S., Koupparis, M. and Georgakopoulos, C., "Gas chromatographic quantitative structure-retention relationships of trimethylsilylated anabolic androgenic steroids by multiple linear regression and partial least squares," J. Chromatogr. A **1216**(47), 8404–8420 (2009).
[14] Chi, Y., Shi, H., Zheng, W. and Sun, J., "Simulating spatial distribution of coastal soil carbon content using a comprehensive land surface factor system based on remote sensing," Sci. Total Environ. **628–629**(6), 384–399 (2018).
[15] Williams, L. J., Abdi, H. and Williams, L. J., "Partial Least Squares Methods: Partial Least Squares Correlation and Partial Least Square Regression," [Computational Toxicology: Volume II], B. Reisfeld and A. N. Mayeno, Eds., Humana Press, Totowa, NJ, 549–579 (2013).
[16] Zalakeviciute, R., Rybarczyk, Y., López-Villada, J. and Diaz Suarez, M. V., "Quantifying decade-long effects of fuel and traffic regulations on urban ambient PM 2.5 pollution in a mid-size South American city," Atmos. Pollut. Res. **9**(1), 66–75 (2018).
[17] Travelagents.com., "Quito City Information and Travel Guide - Quito.com," Travelagents.com, <http://www.quito.com/v/city-info/> (3 August 2017).
[18] U.S. Geological Survey., "EarthExplorer Help Documentation" (2013).
[19] Olmanson, L. G., Brezonik, P. L., Finlay, J. C. and Bauer, M. E., "Comparison of Landsat 8 and Landsat 7 for regional measurements of CDOM and water clarity in lakes," Remote Sens. Environ. **185**, 119–128 (2016).
[20] USGS., "Landsat 8 Surface Reflectance (Provisional) Product Guide : EarthExplorer Version" (2015).
[21] Solano, R., Didan, K., Jacobson, A. and Huete, A., "MODIS Vegetation Index User ' s Guide (MOD13 Series)" (2010).
[22] Secretaria del Ambiente de Quito., "Red Metropolitana de Monitoreo Atmosférico de Quito," 2018, <http://www.quitoambiente.gob.ec/ambiente/index.php/generalidades> (26 June 2018).
[23] ESRI., "Extract Multi Values to Points—Help | ArcGIS Desktop," ArcGIS 10.5, 2017, <http://desktop.arcgis.com/en/arcmap/10.5/tools/spatial-analyst-toolbox/extract-multi-values-to-points.htm> (28 June 2018).
[24] Othman, N., Jafri, M. Z. M. and San, L. H., "Estimating particulate matter concentration over arid region using satellite remote sensing: A case study in Makkah, Saudi Arabia," Mod. Appl. Sci. **4**(11), 131 (2010).
[25] Bilguunmaa, M., Batbayar, J. and Tuya, S., "Estimation of PM10 concentration using satellite data in Ulaanbaatar City," SPIE Asia Pacific Remote Sens.(July), 92591O--92591O (2014).
[26] Ángel, M. and Gutiérrez, R., "Uso de Modelos Lineales Generalizados (MLG) para la interpolación espacial de PM10 utilizando imágenes satelitales Landsat para la ciudad de Bogotá , Colombia," 105–121 (2017).
[27] Khor, W. Y., Hee, W. S., Tan, F., Lim, H. S., Jafri, M. Z. M. and Holben, B., "Comparison of Aerosol optical depth (AOD) derived from AERONET sunphotometer and Lidar system," IOP Conf. Ser. Earth Environ. Sci. **20**(1) (2014).

[28] NASA., "Aerosol Optical Depth," March 2000, <https://earthobservatory.nasa.gov/GlobalMaps/view.php?d1=MODAL2_M_AER_OD> (28 June 2018).
[29] U.S. Geological Survey., "Landsat—Earth observation satellites," Reston, VA (2015).
[30] U.S. Geological Survey., "Product Guide: Landsat 4-7 Surface Reflectance Product" (2018).
[31] U.S. Geological Survey., "Product Guide: Landsat 8 Surface Reflectance Product" (2018).
[32] Vermote, E. F. and Vermeulen, A., "Algorithm Technical Background Document ATMOSPHERIC CORRECTION ALGORITHM : SPECTRAL REFLECTANCES (MOD09) NASA contract NAS5-96062" (1999).
[33] Roger, P. J. C., Vermote, E. F. and Ray, J. P., "MODIS Surface Reflectance User ' s Guide" (2015).
[34] Tucker, C. J., "Red and photographic infrared linear combinations for monitoring vegetation," Remote Sens. Environ. **8**(2), 127–150 (1979).
[35] Janhäll, S., "Review on urban vegetation and particle air pollution - Deposition and dispersion," Atmos. Environ. **105**, 130–137 (2015).
[36] Didan, K., "MOD13Q1 MODIS/Terra Vegetation Indices 16-Day L3 Global 250m SIN Grid V006" (2015).
[37] Wolf, A. F., "Using WorldView-2 Vis-NIR multispectral imagery to support land mapping and feature extraction using normalized difference index ratios," 2012, 83900N–83900N–8.
[38] Lee, J. H., Ryu, J. E., Chung, H. I., Choi, Y. Y., Jeon, S. W. and Kim, S. H., "Development of spatial scaling technique of forest health sample point information," Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci. - ISPRS Arch. **42**(3), 751–756 (2018).
[39] He, J., Zha, Y., Zhang, J. and Gao, J., "Aerosol indices derived from MODIS data for indicating aerosol-induced air pollution," Remote Sens. **6**(2), 1587–1604 (2014).
[40] Ndossi, M. I. and Avdan, U., "Application of open source coding technologies in the production of Land Surface Temperature (LST) maps from Landsat: A PyQGIS plugin," Remote Sens. **8**(5) (2016).
[41] Ghaleb, F., Mario, M. and Sandra, A., "Regional Landsat-Based Drought Monitoring from 1982 to 2014," Climate **3**(3), 563–577 (2015).
[42] Gillespie, A., "Land surface emissivity," [Encyclopedia of Remote Sensing], E. Njoku, Ed., Springer-Verlag New York, 939 (2014).
[43] Vieira, D., Teodoro, A. and Gomes, A., "Analysing Land Surface Temperature variations during Fogo Island (Cape Verde) 2014-2015 eruption with Landsat 8 images," Proc. SPIE **10005**(October), 1000508 (2016).
[44] Sobrino, J. A., Jiménez-Muñoz, J. C., Sòria, G., Romaguera, M., Guanter, L., Moreno, J., Plaza, A. and Martínez, P., "Land surface emissivity retrieval from different VNIR and TIR sensors," IEEE Trans. Geosci. Remote Sens. **46**(2), 316–327 (2008).
[45] Wan, Z., "MODIS Land Surface Temperature Products Users ' Guide" (2013).
[46] Amanollahi, J., Tzanis, C., Abdullah, A. M., Ramli, M. F. and Pirasteh, S., "Development of the models to estimate particulate matter from thermal infrared band of Landsat Enhanced Thematic Mapper," Int. J. Environ. Sci. Technol. **10**(6), 1245–1254 (2013).
[47] Li, Y., Wang, J., Chen, C., Chen, Y. and Li, J., "Estimating PM2.5 in the Beijing-tianjin-hebei region using modis aod products from 2014 to 2015," Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci. - ISPRS Arch. **41**(July), 721–727 (2016).
[48] "Using aerosol optical thickness to predict ground-level PM2.5 concentrations in the St. Louis area: A comparison between MISR and MODIS.", Remote Sens. Environ. **107**(1–2), 33–44 (2007).
[49] Braun, D., Damm, A., Hein, L., Petchey, O. L. and Schaepman, M. E., "Spatio-temporal trends and trade-offs in ecosystem services: An Earth observation based assessment for Switzerland between 2004 and 2014," Ecol. Indic. **89**, 828–839 (2018).
[50] Hossain, M. S., Bujang, J. S., Zakaria, M. H. and Hashim, M., "Assessment of the impact of Landsat 7 Scan Line Corrector data gaps on Sungai Pulai Estuary seagrass mapping," Appl. Geomatics **7**(3), 189–202 (2015).

Modeling the prevalence of respiratory chronic diseases risk using satellite images and environmental data

Cesar I. Alvarez-Mendoza, Ana Teodoro, Juan Ordonez, Andres Benitez, Alberto Freitas, et al.

Cesar I. Alvarez-Mendoza, Ana Teodoro, Juan Ordonez, Andres Benitez, Alberto Freitas, Joao Fonseca, "Modeling the prevalence of respiratory chronic diseases risk using satellite images and environmental data," Proc. SPIE 11157, Remote Sensing Technologies and Applications in Urban Environments IV, 1115705 (2 October 2019); doi: 10.1117/12.2532508

SPIE.

Event: SPIE Remote Sensing, 2019, Strasbourg, France

Modeling the prevalence of respiratory chronic diseases risk using satellite images and environmental data

Cesar I. Alvarez-Mendoza[1,2*], Ana Teodoro[1,3], Juan Ordonez[2], Andres Benitez[2], Alberto Freitas[4] and Joao Fonseca[4]

1 Department of Geosciences, Environment and Land Planning, Faculty of Sciences, University of Porto, 4169-007 Porto, Portugal
2 Universidad Politécnica Salesiana, Grupo de Investigación Ambiental en el Desarrollo Sustentable GIADES, Carrera de Ingeniería Ambiental, Quito, Ecuador.
3 Earth Sciences Institute (ICT), Pole of the FCUP, University of Porto, Portugal.
4 Center for Research in Health Technologies and Information Systems, Faculty of Medicine, University of Porto, Rua Dr. Plácido da Costa, 4200-450 Porto, Portugal.
*****Correspondence: calvarezm@ups.edu.ec; Tel.: +593-9-84647745

ABSTRACT

Several studies have demonstrated that air quality and weather changes have influence in the prevalence of chronic respiratory diseases. Considering this context, the spatial risk modeling along the cities can help public health programs in finding solutions to reduce the frequency of respiratory diseases. With the aim to have a regional coverage and not only data in specific (point) locations, an effective alternative is the use of remote sensing data combined with field air quality data and meteorological data. During the last years, the use of remote sensing data allowed the construction of models to determine air quality data with satisfactory results. Some models using remote sensing based air quality data presented good levels of correlation ($R^2 > 0.5$), proving that it is possible to establish a relationship between remote sensing data and air quality data.

In order to establish a spatial health respiratory risk model for Quito, Ecuador, an empirical model was computed considering data between 2013 and 2017, using the median data values in each parish of the city. The variables are: i) 46 Landsat-8 satellite images with less than 10% of cloud cover and some indexes (normalized difference vegetation index NDVI, Soil-adjusted Vegetation Index SAVI, etc.); ii) air quality data (nitrogen dioxide - NO_2, Ozone - O_3, particulate matter less than 2.5µm - PM2.5 and sulfur dioxide - SO_2) obtained from local air quality network stations and; iii) the hospital discharge rates from chronic respiratory diseases (CRD). In order to establish a probability model to get a CRD, a logistic regression was used. The empirical model is expressed as the probability of occurrence during the studied time. All the procedures were implemented in R Studio. The methodology proposed in this work can be used by health and governmental entities to access the risk of getting a respiratory disease, considering an application of remote sensing in the environmental and health management programs.

Keywords: Landsat-8, Quito, Air quality, health respiratory risk, logistic regression model

1. INTRODUCTION

According to the World Health Organization (WHO), more than 3 million of people have died every year by a chronic respiratory disease (CRD). The CRDs deaths represent approximately 6% of global annual deceases[1]. The CRDs are diseases of the airways where the most common are asthma, chronic obstructive pulmonary disease (COPD), among others. The principal risk factors are the tobacco smoke, air pollution in the cities, occupational chemicals and dust, and frequent lower respiratory infections during childhood[2]. Regarding this, the study of environmental parameters is important considering the direct and indirect relationship between the climate, the environment and the respiratory health[3]. Thus, one of the alternatives to obtain environmental and climate variables is considering remote sensing (RS) data. These data can provide information related to vegetation, urban land use, temperature, retrieve air pollutants and others[4–6]. Regarding this, several studies show an increment in the use of RS in health studies[7,8]. These studies involve infectious disease epidemics and others CRDs, as asthma[9,10].

In the case of use RS data, the most common satellite are from Landsat program and Terra/Aqua Moderate Resolution Imaging Spectroradiometer (MODIS), considering that they are free and easy to download. Typically, the use of RS is related to parameters of vegetation, soil use and climate. Some of the most used indexes are: normalized vegetation

Remote Sensing Technologies and Applications in Urban Environments IV, edited by Thilo Erbertseder, Nektarios Chrysoulakis, Ying Zhang, Proc. of SPIE Vol. 11157, 1115705 · © 2019 SPIE
CCC code: 0277-786X/19/$21 · doi: 10.1117/12.2532508

difference index (NVDI), enhanced vegetation index (EVI), soil-adjusted vegetation index (SAVI), land surface temperature (LST) and others[11–14]. On the other hand, the air pollution has a big influence into the probability to get a CRD. The most common air pollutants are measured in the cities by an automatic air quality network (AQMN). The AQMN measures particulate matter (PM), tropospheric ozone (O_3), sulfur dioxide (SO_2), nitrogen dioxide (NO_2) and others. They are implemented in order to establish a monitoring system in the cities, considering that these air pollutants have a high influence in the incidence in some CRDs and other diseases [15–18]. In order to establish a model to relate health data and other variables, a logistic regression is implemented, considering environmental variables are part of the predictor variables[19,20]. The aim is to find what are the most common independent variables that have the highest probability to be related to CRDs.

In this preliminary study, we compute a model to estimate CRDs in Quito, Ecuador. In order to build the model, we use RS, environmental data and hospital discharge rates (HDR) from CRDs as entry with a logistic regression. The result shows us probability maps from the logistic regression model, where people can see if their zones have or not have a big probability in base to the built model. The idea is to find new alternatives to use RS data in different regions in order to have additional possible health related answers.

2. METHODOLOGY

2.1 Study area

The study area of this work is the city of Quito, Ecuador. The project area is focused on the urban zone, under the influence of AQMN stations and where most of the people live. In the study area, the considered zones are the parishes or "parroquias" as the unit, because of the availability of hospital discharge rates only at this level of information (Figure 1). Quito is located in the middle of Andean region with a middle latitude, having a constant temperature during all the year. The mean Quito altitude is about 2800m amsl.

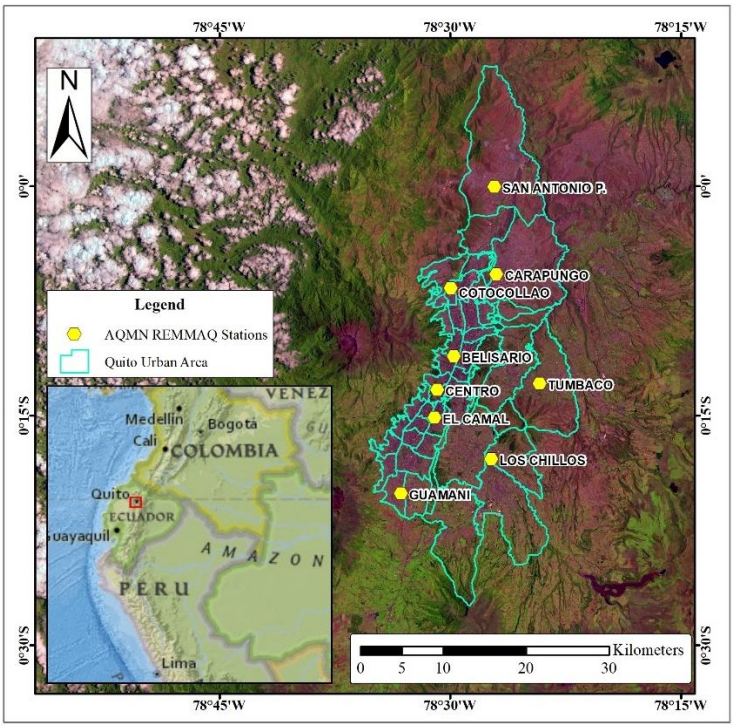


Figure 1. Study location (Quito, Ecuador). The green polygons are the parishes or "parroquias" and the yellow points are the AQMN stations. The base image is a Landsat-8 OLI from 20/09/2017.

2.2 Methodology

The challenge of the study is to compute a risk for CRDs (predictive model), considering the RS and environmental data (air pollution) as input in the model (independent variables) between 2013 to 2017. Thus, a final table where are matched the RS, the air pollution parameters and the hospital discharge rates data (Table 1) is established, with the objective to build a final model and show the spatial distribution. The general methodology is showed in the Figure 3.

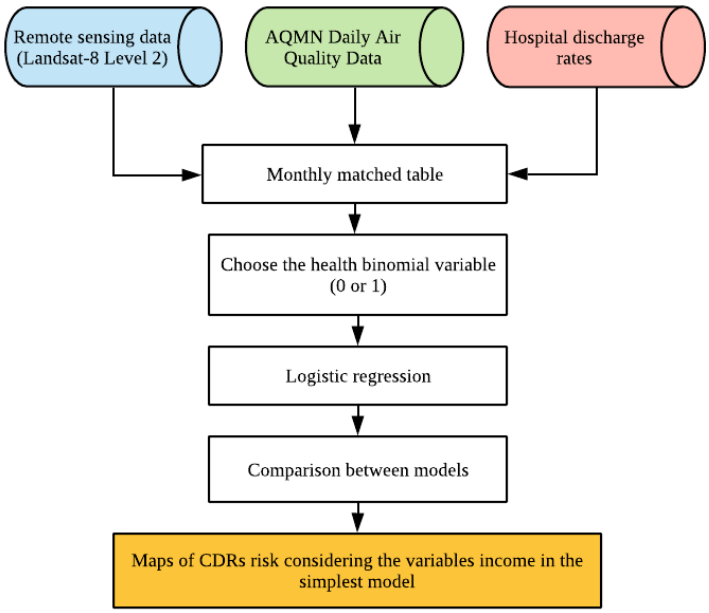


Figure 3. Summary methodology workflow.

Table 1. Input model variables

Data type	Variable	Units
Remote Sensing	Band 1 – Coastal aerosol (B1) Band 2 - Blue (B2) Band 3 – Green (B3) Band 4 – Red (B4) Band 5 – Near Infrared NIR (B5) Band 6 – Short-wave infrared SWIR 1 (B6) Band 7 – Short-wave infrared SWIR 2 (B7) NDVI SAVI EVI LST	Surface reflectance (Landsat 8 – Level 2)
Environmental data - Air pollution	PM2.5 SO_2 O_3 NO_2	$\mu g/m^3$
Hospital discharge rates	CRDs admission rate (per 10,000) (NS)	Hospital discharge rates divided per population

2.3 Extracting remote sensing variables

Three classes of variables are chosen in this study (RS, AQMN data and HDR) between 2013 to 2017. Considering the RS data, forty six Landsat-8 Level-2[21] images are used in the study. These images are geometric, radiometric and atmospheric corrected[22]. They were downloaded from Land Satellites Data Systems Science Research and Development (LSRD) (https://espa.cr.usgs.gov/). The available download products are: (i) land surface reflectance bands from the Operational Land Imager (OLI) sensor; (ii) some environmental indexes already computed (NDVI, SAVI, EVI) and; (iii) the brightness temperature (BT). Moreover, we contemplate only images with less than ten percentage of cloud cover in the study area, considering that Quito is a city with a high cloud density during all the year, and so some methods to remove clouds have been applied in order to recover some image data[13,23].

The RS variables used as predictor are the surface reflectance and some environmental indexes. In the case of the surface reflectance, the variables are B1, B2, B3, B4, B5, B6 and B7. The reason to consider them is the high relationship between the OLI bands and air pollution[5,24–27]. On the other hand, we use some environmental indexes as NDVI, SAVI and EVI. They are used because they have a high relationship with the vegetation type and coverage and the land use. Moreover, in order to have a climate variable, the LST is computed. It is computed from the BT by Inversion of Planck's function[28] as presented in the Equation 1.

$$LST = \frac{BT}{\left(1+\left(\frac{\lambda * BT}{\rho}\right) ln\varepsilon\right)} \tag{1}$$

Where, λ is the center wavelength (Landsat-8 = 10.8 μm)[29], ρ is the a constant (Equation 2) and ε is the emissivity (Equation 3).

$$\rho = \frac{h * c}{s} \tag{2}$$

Where, h is the Planck's constant (6.626e-14 Js), c is the velocity of light (2.998e-8 m/s) and s is the Boltzmann constant (6.626 e-34 J/K).

The emissivity (ε) is the efficiency with a surface emits heat as Thermal Infrared (TIR) radiation[30]. The Equation 3 is a semi-empirical algorithm where the variations of NDVI in the vegetation ($NDVI_v$) and soil ($NDVI_s$) are considered[31].

$$\varepsilon = \begin{cases} \varepsilon_s, \ NDVI < NDVI_s \\ \varepsilon_s + (\varepsilon_v - \varepsilon_s) P_V, \ NDVI_s \leq NDVI \leq NDVI_v \\ \varepsilon_v, \ NDVI > NDVI_v \end{cases} \tag{3}$$

Where, ε_s is the emissivity for soil (0.973). ε_v is the emissivity for vegetation (0.985)[32]. P_v is the proportion of vegetation in the study area (Equation 4), where $NDVI_v$ and $NDVI_s$ are 0.2 and 0.5[31].

$$P_V = \left(\frac{NDVI - NDVI_s}{NDVI_v - NDVI_s}\right)^2 \tag{4}$$

With the forty-six Landsat-8 Level 2 images, we extract a median pixel value in each parish every month from the RS variables. The extraction considers a previous analysis, where the pixels with clouds are not considered in order to obtain the median value. This is done considering the Landsat-8 thin cloud band (B9). All the extraction and computation process were performed with R software. The final output is a medium RS data monthly table.

2.4 Air Quality Measurements

On the other hand, the collected data from AQMN are air daily measures. The AQMN in Quito is the "Red Metropolitana de Monitoreo Atmosférico de Quito" (REMMAQ)[33]. It is working since 2002 and it is composed by nine monitoring stations in some city zones (Figure 1). REMMAQ stations get meteorological and air pollution variables. In our case, we consider the main air pollutions related to respiratory health. They are particulate matter less than 2.5 microns (PM2.5), SO_2, O_3 and NO_2. The data is available in "Secretaria del Ambiente" page (http://www.quitoambiente.gob.ec/ambiente/index.php/datos-horarios-historicos). All the process to normalize the data was realized on R software, where we choose a median value between each month and parish in order to match with RS and HDR data in a unique input data table. The inverse distance weighted (IDW) method was applied to build the raster data to extract each variable by parish.

2.5 Hospital discharge rates (HDR)

Finally, the CRDs admission rates are obtained from "Instituto Nacional de Estadisticas y Censos" (INEC). INEC provides this data from 2012 to 2017 in each parish, based in the International Classification of Diseases 10 (ICD-10) from the WHO[34]. The HDR data can be downloaded from INEC web page (http://www.ecuadorencifras.gob.ec/camas-y-egresos-hospitalarios/). In order to have only the CRDs admission rates, we filter the INEC tables selecting the cases of chronic lower respiratory diseases (ICD-10 codes: J40 – J47) per month and parish. Furthermore, we also collect the population data from INEC in order to obtain the HDR per 10000 people. With this new computed variable, we generate the binomial dependent variable to be modeled. This was done considering a cutoff in a main break value in the analysis of the histogram. The final HDR data is showed in a binomial variable (0 or 1) per month, year and parish.

2.6 CRDs risk modeling

In order to compute the model, the technique employed was the multiple logistic regression. This method needs a binomial variable as dependent variable. Several health studies use the logistic regression to established probability models[35], considering some predictors to analyze if they have or not relationship with the binomial dependent variable (0 or 1), which is a classification variable.

The Equation 5 shows the model considering all variables.

$$PS = \frac{1}{1+e^{-(I+a*B1 + b*B2 + c*B3 + d*B4 + e*B5 + f*B6 + g*B7 + h*NDVI + i*SAVI + j*EVI + k*LST + l*NO2 + m*O3 + n*PM2.5 + o*SO2)}} \quad (5)$$

Where, a,b,…,o are coefficients computed from the multiple logistic regression from the independent variables, I is the intercept and PS is the probability to have or not a CDRs.

Nevertheless, the objective is to build the simplest model with few predictors, thus, the backward stepwise selection method was applied to obtain the model with less variables through the lowest Akaike information criterion (AIC). Moreover, the final process is to analyse if the model with the lowest AIC has correlation variables, specifically the RS variables. If they have correlation, only a variable is selected between them in order to establish the final model and to elaborate the risk maps.

3. RESULTS AND DISCUSSION

The final monthly parish data table is the result of combining 892 observations from RS, air pollution and HDR data between 2013 to 2017. Considering the requirement to have a binomial dependent variable, the HDR histogram was analyzed (Figure 4) in order to define a cutoff value. A cutoff value of 0.35 HDR per 10000 people was them selected. It means if the HDR per 10000 people have a value less than 0.35, it takes the value 0 and, if the value is more than 0.35, it takes a value 1 (Table 2). In this aspect, we consider a parish with less than 0.35 HDR per 10000 people without sick people.

Thus, the model is built considering a multiple linear regression, where the backward stepwise selection gives a new model with the lowest AIC value. The Equation 6 shows the new model with 8 independent variables as predictors, where most of them are RS variables related to vegetation and soil use.

$$P(Y = 1) = \frac{1}{1+e^{-(I+a*B1 + b*B2+d*B4+f*B6+ h*NDVI + i*SAVI + j*EVI+ o*SO2)}} \quad (6)$$

On the other hand, considering the evaluation of multicollinearity in Equation 6, where some of the predictors have a high correlation value (near to 1), they are discarded in the final model[36]. For example, the B1, B2 and B4 are the same variables according to the correlation graphic. In this case, we used only a variable (B2) considering the relationship of blue band with humidity and it possible relationship with CRDs[37]. The rest of them were discarded in the final model. Another group of variables to discard is between NDVI, SAVI and EVI, where NDVI was selected according to the importance of this variable in most of the studies.

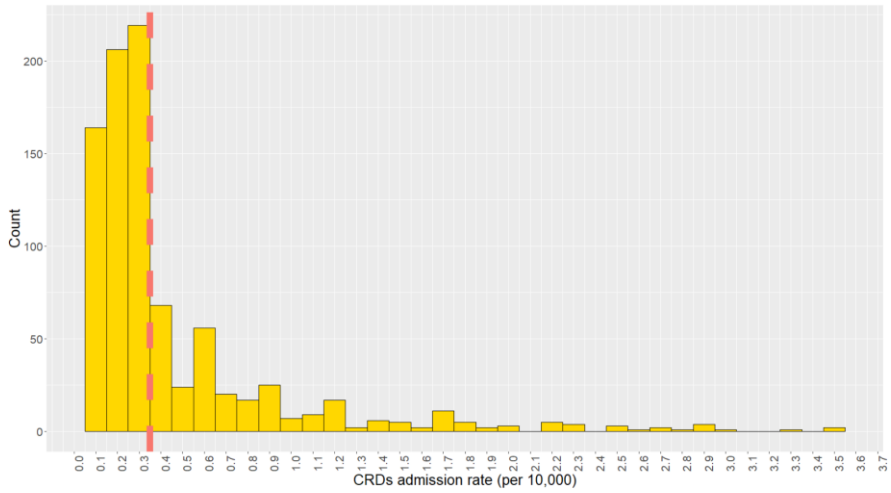


Figure 4. Histogram of CRDs admission rate per 10000 people. The red line is the cutoff value to define the binomial variable (0.35).

Table 2. Definition of new binomial variable considering a cutoff value

Category – Binomial value	Number of observations	% of observations
0 (NS < 0.35)	589	66
1 (NS > 0.35)	303	44
Total	892	100

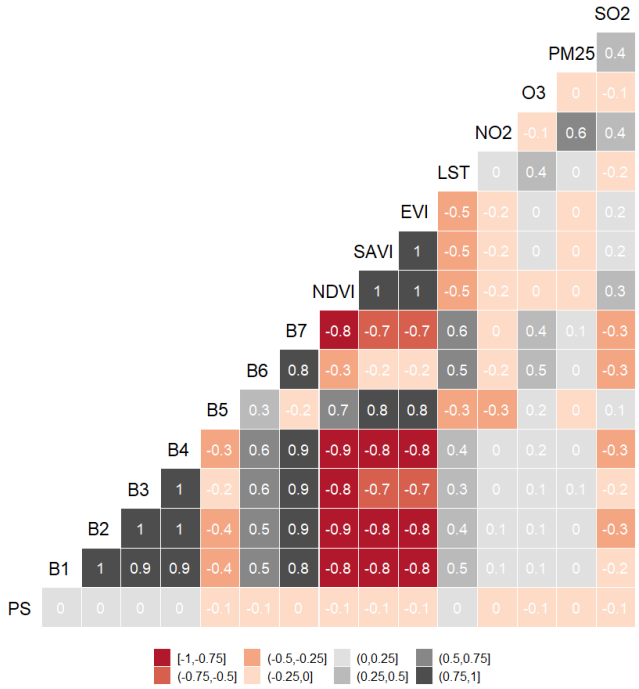


Figure 5. Correlation graphic. Most of the RS variables have a high correlation value.

The Equation 7 presents the final model considering the correlation analysis (Figure 5). The final predictors are B2, B6, NDVI and SO2, meaning that only three RS variables and one air pollution variable entry in the final model.

$$P(Y=1) = \frac{1}{1+e^{-(\mathrm{I}+\mathrm{b}*\mathrm{B2}+\mathrm{f}*\mathrm{B6}+\mathrm{h}*\mathrm{NDVI}+\mathrm{o}*\mathrm{SO2})}} \quad (7)$$

In the evaluation of parameters (Table 3), we see that the more significative variables are SO_2 and B6, meaning that the probability to get a CRDs with this data are in areas with a high response of the short-wave infrared. Some authors relate the infrared with the presence of O_3[5,38] and O_3 with the presence of CRDs[16,19]. On the other hand, the SO_2 is related with asthma in some recent studies as a risk factor[39,40].

Table 3. Final model parameters

Variable	Coefficients - Estimate	Significance	Odds ratio (OR)
Intercept (I)	2.53030	0.006	12.557
B2	b = -6.04123	0.601	0.0024
B6	f = -7.62867	0.061	0.0004
NDVI	h = -1.42922	0.219	0.2395
SO2	o = -0.21053	0.000	0.8101

The final model is evaluated in a relative operating characteristic (ROC) curve with an area under the curve (AUC) of 0.609. This suggests a probability of 61% of correctly classifying between the two classes (having CRDs or not having a CRDs) [41,42].

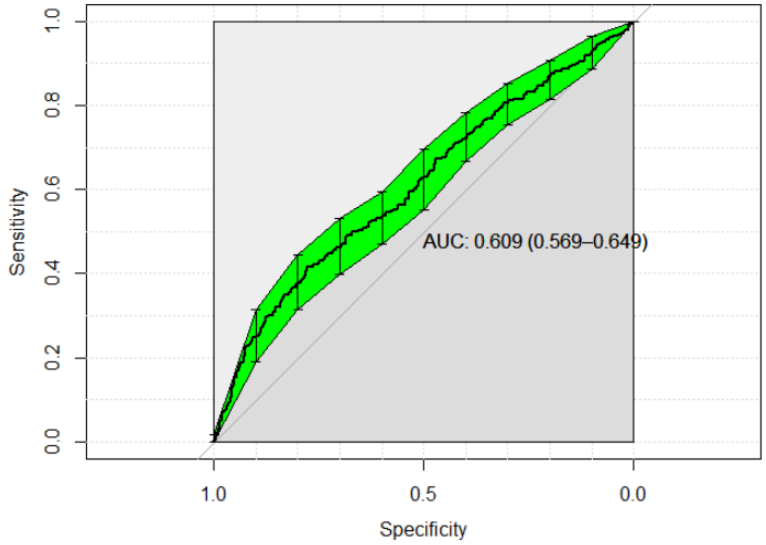


Figure 6. AUC of the final model

Finally, the logistic model is represented over monthly maps in order to compare what can be the risk to get CRDs according to color levels (Figure 7), where red is a high risk and blue low risk. In this case, the final maps are created with a spatial resolution of 30 meters, considering the Landsat-8 bands and environmental indexes computed by IDW. Thus, we have maps with a medium resolution in order to know the probability to get a CRD with more accuracy if it is combined with other spatial information as roads, hospital locations, etc. For example, on Figure 7 maps, we have more probability to get CRDs in regions with more road density and less probability in regions with more green area.

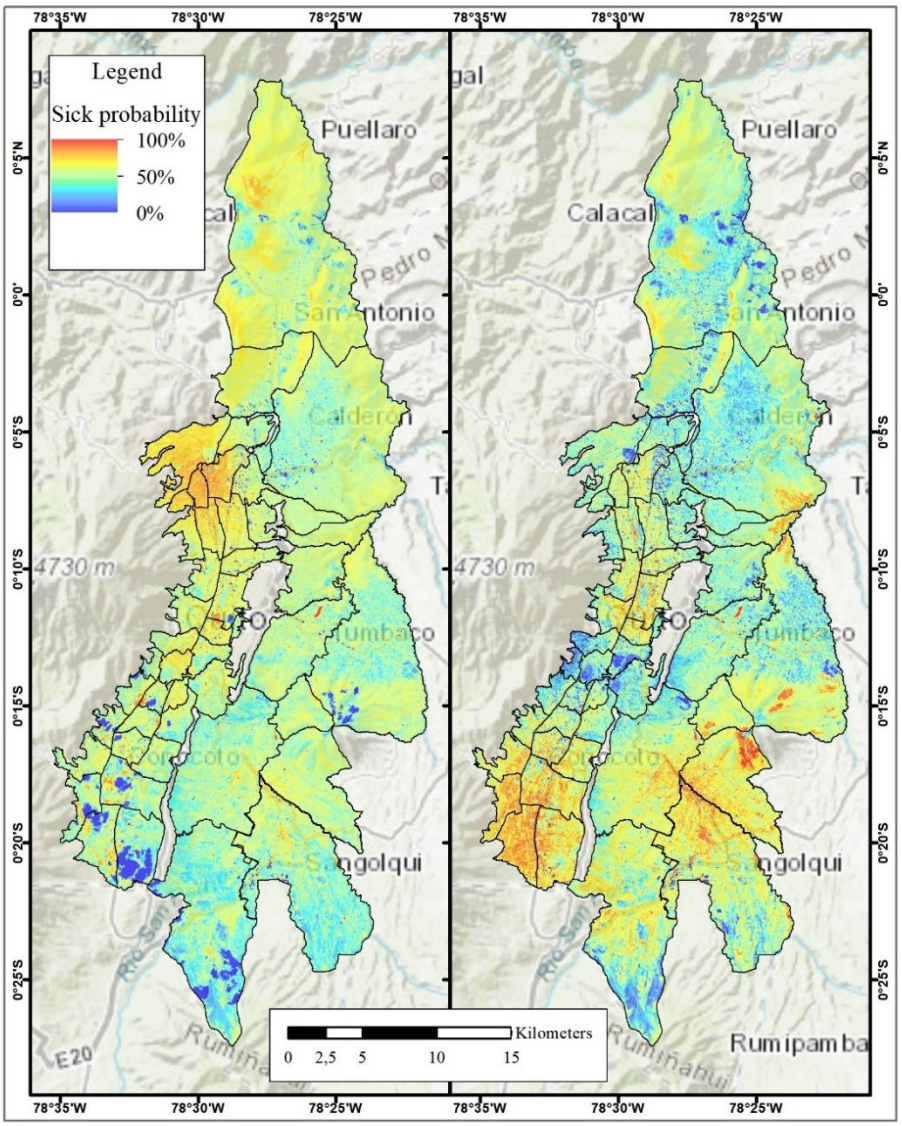


Figure 7. Probability maps to get CRDs in base to the final model computed. On the left the map in September 2013. On the right, the map in September 2015.

4. CONCLUSIONS

This preliminary study investigates a possible relationship between remote sensing, environmental variables and hospital discharge rates by chronic respiratory diseases in Quito, Ecuador. The model established uses a multiple logistic regression considering parishes where it is possible the presence of a sick people by a chronic respiratory disease. The results show a model with four variables; where three of them were obtained by remote sensing and one by air quality measures. The most significant variables are the short-wave infrared or band 6 in Landsat-8 and sulfur dioxide (SO_2). Moreover, the AUC of the model was 0.609. Considering this model evaluation, we generated risk maps. This kind of models can be an interesting alternative tool to health authorities in order to evaluate the public health with remote sensing variables.

REFERENCES

[1] WHO., "Chronic respiratory diseases (CRDs)," WHO, 2019, <https://www.who.int/respiratory/en/> (21 June 2019).

[2] O'Connor, G. T., Neas, L., Vaughn, B., Kattan, M., Mitchell, H., Crain, E. F., Evans, R., Gruchalla, R., Morgan, W., Stout, J., Adams, G. K. and Lippmann, M., "Acute respiratory health effects of air pollution on children with asthma in US inner cities," J. Allergy Clin. Immunol. **121**(5), 1133–1139.e1 (2008).

[3] Barry, M. and Annesi-Maesano, I., "Ten principles for climate, environment and respiratory health," Eur. Respir. J. **50**(6), 1701912 (2017).

[4] Lee, J. H., Ryu, J. E., Chung, H. I., Choi, Y. Y., Jeon, S. W. and Kim, S. H., "Development of spatial scaling technique of forest health sample point information," Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci. - ISPRS Arch. **42**(3), 751–756 (2018).

[5] Alvarez-Mendoza, C. I., Teodoro, A. and Ramirez-Cando, L., "Spatial estimation of surface ozone concentrations in Quito Ecuador with remote sensing data, air pollution measurements and meteorological variables," Environ. Monit. Assess. **191**(3), 155 (2019).

[6] Alvarez-Mendoza, C. I., Teodoro, A., Torres, N., Vivanco, V. and Ramirez-Cando, L., "Comparison of satellite remote sensing data in the retrieve of PM10 air pollutant over Quito, Ecuador," Remote Sens. Technol. Appl. Urban Environ. III **10793**, Y. Z. Thilo Erbertseder, Nektarios Chrysoulakis, Ed., 107930I–10793–12, Berlin (2018).

[7] Viana, J., Vasco Santos, J., Manuel Neiva, R., Souza, J., Duarte, L., Cláudia Teodoro, A. and Freitas, A., "Remote Sensing in Human Health: A 10-Year Bibliometric Analysis," Remote Sens. (2017).

[8] Seltenrich, N., "Remote-sensing applications for environmental health research.," Environ. Health Perspect. **122**(10), A268-75 (2014).

[9] Samson, D. M., Archer, R. S., Alimi, T. O., Arheart, K. L., Impoinvil, D. E., Oscar, R., Fuller, D. O. and Qualls, W. A., "New baseline environmental assessment of mosquito ecology in northern Haiti during increased urbanization," J. Vector Ecol. **40**(1), 46–58 (2015).

[10] Ayres-Sampaio, D., Teodoro, A. C., Sillero, N., Santos, C., Fonseca, J. and Freitas, A., "An investigation of the environmental determinants of asthma hospitalizations: An applied spatial approach," Appl. Geogr. **47**, 10–19 (2014).

[11] Wang, Y., Du, H., Xu, Y., Lu, D., Wang, X. and Guo, Z., "Temporal and spatial variation relationship and influence factors on surface urban heat island and ozone pollution in the Yangtze River Delta, China," Sci. Total Environ. **631**–**632**, 921–933 (2018).

[12] Herring, J. W. and D., "Measuring Vegetation (NDVI & EVI)," August 2000.

[13] Alvarez-Mendoza, C. I., Teodoro, A. and Ramirez-Cando, L., "Improving NDVI by removing cirrus clouds with optical remote sensing data from Landsat-8 – A case study in Quito, Ecuador," Remote Sens. Appl. Soc. Environ. (2019).

[14] Chejarla, V. R., Maheshuni, P. K. and Mandla, V. R., "Quantification of LST and CO<inf>2</inf> levels using Landsat-8 thermal bands on urban environment," Geocarto Int. **31**(8) (2016).

[15] Sweileh, W. M., Al-Jabi, S. W., Zyoud, S. H. and Sawalha, A. F., "Outdoor air pollution and respiratory health: A bibliometric analysis of publications in peer-reviewed journals (1900 - 2017)," Multidiscip. Respir. Med. **13**(1), 1–12 (2018).

[16] US EPA, O., "Health Effects of Ozone Pollution," <https://www.epa.gov/ozone-pollution/health-effects-ozone-pollution> (24 May 2018).

[17] Kallawicha, K., Chuang, Y., Lung, S. C., Wu, C., Han, B., Ting, Y. and Chao, H. J., "Outpatient Visits for Allergic Diseases are Associated with Exposure to Ambient Fungal Spores in the Greater Taipei Area," 2077–2085 (2018).

[18] Guarnieri, M. and Balmes, J. R., "Outdoor air pollution and asthma.," Lancet (London, England) **383**(9928), 1581–1592 (2014).

[19] O' Lenick, C. R., Chang, H. H., Kramer, M. R., Winquist, A., Mulholland, J. A., Friberg, M. D. and Sarnat, S. E., "Ozone and childhood respiratory disease in three US cities: evaluation of effect measure modification by neighborhood socioeconomic status using a Bayesian hierarchical approach," Environ. Heal. **16**(1), 36 (2017).

[20] Rawi, N. A. M. N., Jalaludin, J. and Chua, P. C., "Indoor Air Quality and Respiratory Health among Malay Preschool Children in Selangor.," Biomed Res. Int. **2015**, 248178 (2015).

[21] Department of the Interior U.S. Geological Survey., "Landsat 8 (L8) Data Users Handbook" (2016).

[22] U.S. Geological Survey., "Product Guide: Landsat 8 Surface Reflectance Product" (2018).
[23] Alvarez, C. I., Teodoro, A. and Tierra, A., "Evaluation of automatic cloud removal method for high elevation areas in Landsat 8 OLI images to improve environmental indexes computation," Proc. SPIE 10428, Earth Resour. Environ. Remote Sensing/GIS Appl. VIII 1042809 **10428**, SPIE, Ed., 1042809–1042812, Warsaw (2017).
[24] Othman, N., Jafri, M. Z. M. and San, L. H., "Estimating particulate matter concentration over arid region using satellite remote sensing: A case study in Makkah, Saudi Arabia," Mod. Appl. Sci. **4**(11), 131 (2010).
[25] Bilguunmaa, M., Batbayar, J. and Tuya, S., "Estimation of PM10 concentration using satellite data in Ulaanbaatar City," SPIE Asia Pacific Remote Sens.(July), 92591O--92591O (2014).
[26] Alvarez, C. I. and Padilla Almeida, O., "Estimación de la contaminación del aire por PM10 en Quito a través de índices ambientales con imágenes LANDSAT ETM+.," Rev. Cart.(92), 135–147 (2016).
[27] Ángel, M. and Gutiérrez, R., "Uso de Modelos Lineales Generalizados (MLG) para la interpolación espacial de PM10 utilizando imágenes satelitales Landsat para la ciudad de Bogotá , Colombia," 105–121 (2017).
[28] Ndossi, M. I. and Avdan, U., "Application of open source coding technologies in the production of Land Surface Temperature (LST) maps from Landsat: A PyQGIS plugin," Remote Sens. **8**(5) (2016).
[29] Ghaleb, F., Mario, M. and Sandra, A., "Regional Landsat-Based Drought Monitoring from 1982 to 2014," Climate **3**(3), 563–577 (2015).
[30] Gillespie, A., "Land surface emissivity," [Encyclopedia of Remote Sensing], E. Njoku, Ed., Springer-Verlag New York, 939 (2014).
[31] Vieira, D., Teodoro, A. and Gomes, A., "Analysing Land Surface Temperature variations during Fogo Island (Cape Verde) 2014-2015 eruption with Landsat 8 images," Proc. SPIE **10005**(October), 1000508 (2016).
[32] Sobrino, J. A., Jiménez-Muñoz, J. C., Sòria, G., Romaguera, M., Guanter, L., Moreno, J., Plaza, A. and Martínez, P., "Land surface emissivity retrieval from different VNIR and TIR sensors," IEEE Trans. Geosci. Remote Sens. **46**(2), 316–327 (2008).
[33] Secretaria del Ambiente de Quito., "Red Metropolitana de Monitoreo Atmosférico de Quito," 2018, <http://www.quitoambiente.gob.ec/ambiente/index.php/generalidades> (26 June 2018).
[34] WHO., "ICD-10 Version:2016," WHO, 2016, <https://icd.who.int/browse10/2016/en> (23 June 2019).
[35] Ghanname, I., Chaker, A., Cherkani Hassani, A., Herrak, L., Arnaul Ebongue, S., Laine, M., Rahhali, K., Zoglat, A., Benitez Rexach, A. M., Ahid, S. and Cherrah, Y., "Factors associated with asthma control: MOSAR study (Multicenter Observational Study of Asthma in Rabat-Morocco)," BMC Pulm. Med. **18**(1), 61 (2018).
[36] Ranganathan, P., Pramesh, C. S. and Aggarwal, R., "Common pitfalls in statistical analysis: Logistic regression.," Perspect. Clin. Res. **8**(3), 148–151 (2017).
[37] Hayes, D., Collins, P. B., Khosravi, M., Lin, R.-L. and Lee, L.-Y., "Bronchoconstriction Triggered by Breathing Hot Humid Air in Patients with Asthma," Am. J. Respir. Crit. Care Med. **185**(11), 1190–1196 (2012).
[38] Petruci, J. F. da S., Fortes, P. R., Kokoric, V., Wilk, A., Raimundo, I. M., Cardoso, A. A. and Mizaikoff, B., "Real-time monitoring of ozone in air using substrate-integrated hollow waveguide mid-infrared sensors," Sci. Rep. **3**, 3174 (2013).
[39] Andersson, E., Knutsson, A., Hagberg, S., Nilsson, T., Karlsson, B., Alfredsson, L. and Torén, K., "Incidence of asthma among workers exposed to sulphur dioxide and other irritant gases," Eur. Respir. J. **27**(4), 720–725 (2006).
[40] Kuo, C.-Y., Chan, C.-K., Wu, C.-Y., Phan, D.-V. and Chan, C.-L., "The Short-Term Effects of Ambient Air Pollutants on Childhood Asthma Hospitalization in Taiwan: A National Study.," Int. J. Environ. Res. Public Health **16**(2) (2019).
[41] Mandrekar, J. N., "Receiver Operating Characteristic Curve in Diagnostic Test Assessment," J. Thorac. Oncol. **5**(9), 1315–1316 (2010).
[42] Mason, S. J. and Graham, N. E., "Areas beneath the relative operating characteristics (ROC) and relative operating levels (ROL) curves: Statistical signii cance and interpretation" (2002).