Multi-Source Data Modelling to Understand the Effects of Tourism Demand on Air Quality in Italy

Adam Kadiev

Multi-Source Data Modelling to Understand the Effects of Tourism Demand on Air Quality in Italy

THESIS

submitted in fulfillment of the
requirements for the degree of

MASTER OF SCIENCE

in

COMPUTER SCIENCE

by

Adam Kadiev

To be defended publicly on Monday June 26, 2023 at 1:00 PM

Web Information Systems Research Group
Department of Software Technology
Faculty EEMCS, Delft University of Technology
Delft, the Netherlands
www.ewi.tudelft.nl

Multi-Source Data Modelling to Understand the Effects of Tourism Demand on Air Quality in Italy

Author: Adam Kadiev
Student id: 4918002
Email: a.a.kadiev@student.tudelft.nl

Abstract

The goal of this research is to model and understand the effects of tourism demand on air quality by performing data integration on multi-source data. This research is aimed at researchers and practitioners aiming to perform multidisciplinary research in the fields of data science and geoscience, presenting the methods and challenges that arise when performing such an analysis. A data processing pipeline explains the research from a data integration perspective involving the data retrieval and pre-processing tasks. This enables the construction of datasets for machine learning modelling and prediction of air pollutant levels based on tourism data. The study area of this research is Italy which is chosen based on its significant tourism industry and wide availability of data about tourism development. For this study, in situ air quality data sampled using Google Earth Engine (GEE) around accommodation, transportation and tourism attraction locations is modelled with tourist arrival numbers, nights spent and average length of stay. Long short-term memory (LSTM) multivariate time series modelling is performed afterwards to understand predictability of air quality on a national and regional level. To this end, this research looks into three different stages of the modelling process of tourism with air quality which are: (i) retrieving accommodation, transportation and tourism attraction locations using the RDF model, (ii) identifying which pollutants are correlated and Granger-caused by the different tourism demand features using sampled satellite air quality data of the identified tourism locations, (iii) understanding performance characteristics of LSTM time series models by training on tourism demand and air quality data. Correlation analysis indicates the potential to model the relation between tourism demand indicators and $PM_{2.5}$ in overall cleaner regions in terms of this pollutant. In these regions, Granger-causality testing suggests a higher chance of predictability of $PM_{2.5}$ time series using tourism demand data from the previous month. Training an LSTM model using the information of this lagged relationship suggests that regions with overall high $PM_{2.5}$ levels are challenging to model showing high RMSE scores. Training an LSTM model for these regions also required more epochs compared to overall cleaner regions to model the effects of tourism demand on air quality.

Thesis Committee:

Chair:	Prof. Dr. Christoph Lofi, Web Information Systems, TU Delft
Committee Member:	Prof. Dr. Soham Chakraborty, Programming Languages, TU Delft
Daily supervisor:	Dr. Lixia Chu, Web Information Systems, TU Delft

Contents

List of Figures

Chapter 1

Introduction

The focus of this research study is to model the effects of tourism demand on air quality and provide the reader with an overview of the data integration and modelling process for this analysis. Researchers and practitioners in the fields of data analysis and geoscience can use this research to gain insight into how multi-source data integration can be performed in these contexts. Specifically, analysing the datasets created through data integration can provide insights into how tourism demand can model air quality levels at different categories of tourism sites given the different air quality conditions prevalent. Correlation and causality testing of the studied variables can be used to work towards training a machine learning model to use tourism demand indicators to model air quality.

By combining data about tourism demand in Italy, satellite data on air quality levels over time and locations of categorised tourism sites, this research models impact on air quality levels using the long short-term memory (LSTM) model. This model is an artificial neural network that is often used for time series analysis. Different industry sectors and geographical circumstances affect air quality levels in ways which can make it challenging for a model to learn the effects of the tourism industry. To better understand this effect, the analysis in this research is performed on two different spatial levels, namely on a national and regional level. The analysis of tourism demand on air quality modelling on a national level is performed by modelling Italy as a whole which covers regions potentially affected by different air quality conditions and other polluting industries besides the tourism industry. By also looking at regions where the confounding effects on air quality due to other industries and overall air quality levels are less prevalent, this research studies how well the national level results on air pollutant measurements relate to the regional level analysis in terms of the correlation between tourism arrivals and nights spent. Further understanding of the relationship between tourism demand indicators and the pollutant levels through correlation, Granger-causality, and multivariate LSTM time series analysis can give insights to better model the effects of tourism demand on air quality. The benefit of modelling the relationship between tourism arrivals and nights spent on an air pollutant on a regional level is that the tourism data is about a smaller spatial region making the data more meaningful to the area that is analysed compared to the data on national level. As the regional analysis also models regions with overall better air quality and less polluting industries like southern regions of Italy, the regional analysis is expected to provide insights on how the modelling

performance of air pollution based on tourism demand varies under different air quality conditions. Regional differences in correlation and Granger-causality between higher and lower polluted regions can give key insights in the how the tourism demand and pollutant time series relate to each other in terms of predictability.

One of the important facilities that drives the tourism industry is transportation and for many people, the use of public transport is the primary method of traveling to the country visited. This form of travel consists of air travel and other ground transportation like train trips and bus rides. Railway station and bus stations often form importation junctions where other car traffic like taxis operate to bring people to their accommodation place and the tourist sites they want to visit. Depending on the demand of tourism, the intensity of use of these facilities increases or decreases potentially effecting air pollutant levels. With the aim of understanding sustainable tourism, it is valuable to analyse the impact of tourism on air quality across different locations important for tourism in order to maintain the quality of the touristic destinations.

Research on the tourism industry indicates that air emissions are positively correlated with economic growth [32]. Linking this with tourist arrivals and their use of transportation could give insight on the impact of transportation on air quality levels for different pollutants as well. Sustainable development around tourism is seen as an important concept as the environment in which it takes place greatly determines its well-being [8]. Air quality levels determine a great amount to shaping a healthy environment and it is therefore important to research the effect of tourism demand on this environmental indicator across different types of places important for tourism.

Finding the locations of categorised tourism sites involves describing the different places of interest using characteristics like the type of accommodation, its coordinates and region. Fortunately, there are web services available making use of the Resource Description Framework (RDF) data model which can be used together with a query language like SPARQL to describe these characteristics or relationships in the context of RDF. The use of predicates allows one to describe relationships between entities in the world through unique URIs for each entity. In the case of finding categorised tourism sites, the use of relationships like describing the country of the site and what location it represents, for instance a hotel or airport, are important methods that semantic web applications like the Wikidata Knowledge Base enable to perform.

The use of machine learning for time series analysis is an important part of this research as it allows to understand how well tourism effect can be modelled in terms air quality predictability over time. Correlation analysis with different types of tourism demand indicators like arrivals on a national level and regional level are aimed at better understanding the relationship between tourism and specific air pollutants that are analysed, while at the same time looking at patterns between tourism and air quality across accommodation, transportation and tourist attractions.

Retrieving air quality data is often done by using air quality stations on the ground. These stations can provide detailed data about the area in which they are located. The downside of ground-based stations to measure air quality is that they can only capture air quality metrics in its surroundings. Fortunately, there are also air quality monitoring systems on satellites that can cover the entire world surface. In this case it is important that the data

the satellite collects is validated as different kinds of atmospheric conditions can make it challenging in some cases to collect reliable air quality data. For this, in situ data is valuable as it combines other types of measured data like ground level measurements together with satellite data to obtain calibrated and verified results [36].

By combining these research objectives, the overarching research question of this work is as follows: *How can data integration be used to enable the modelling of the effects of tourism demand on different air pollutants using a data processing pipeline operating on categorised RDF tourism locations data, tourism statistics and air pollution satellite measurements?*

To guide the process of formulating the answer to this question, the following sub-questions are formulated:

1. What are challenges of a data processing pipeline for integrating data on satellite measurements, tourism demand data and categorised tourism sites?

2. What air pollutants are linked with tourism demand indicators at different categorical tourism sites based on the integrated data in terms of correlation and causality analysis?

3. What are the performance characteristics of the final machine learning time series models trained on a regional and national level using data based on the correlated tourism and air quality indicators?

Chapter 2

Related Work

This chapter discusses previous work that is related to measuring air quality in the context of tourism through different kinds of modelling methods. It focuses on how tourism is defined in these studies and the approaches to measuring air quality together with the tools and technologies that are used for the analysis.

2.1 Modelling techniques of tourism and air quality

Studies suggest that it is important to understand the relationship between the tourism industry and air quality which would give better insight in sustainability practises regarding both the tourism industry and air quality levels [19]. In literature, this relationship is studied in both directions using different variables to capture tourism and air quality [30, 15]. Existing studies that look at the impact of tourism on air quality indicate that there exists an impact of tourism on the environment [28]. These studies use proxy values for analysing the tourism demand retrieved from regional panel data and find that the direction of causality of tourism and air quality impact differs depending on which region is analysed [1, 52]. These kinds of studies have mostly been carried out in East Asia where studies link the rapid economic growth with the increase of tourism. Analysis of the tourism industry is said to be augmenting the energy consumption in national regions leading to increased carbon emissions [3] with similar studies indicating that these carbon emissions from tourism should be used for setting local goals and emission mitigation [47]. The source of these emissions as a result of tourism is primarily accounted to the use of transportation, energy use for tourism housing and delivering of amenities for the tourism industry [33]. Regarding this analysis of tourism impact of air quality levels, literature indicates the potential of modelling air pollution like carbon emissions together with energy consumption and tourism to better understand equilibrium relationships on the long-term regarding these variables [27, 53]. Focusing on carbon emissions, research also indicates that tourism has the potential to lower emissions when it is used to replace high-emission industries which was found though statistical significance testing of different time spans of tourism development and emissions analysis [4]. Similarly, specialised spatial econometric methods can be used to analyse the relationship between the development of a tourism industry with its carbon

emissions [31].

In literature, different methods are used for modelling the air quality levels at a study area. A study that looks at air quality influence by tourism in Mallorca indicated that the use of a General Additive Model (GAM) to model air quality is better able to capture air quality behaviours compared to linear regression models as air quality is unlikely to behave in a linear and additive fashion [16]. A similar study that explores the contribution of tourism on air pollution focusing on tropospheric ozone levels found that models can capture the daily rising tropospheric ozone caused by activities like transportation and air conditioning [41]. The authors of the study also indicate that it is important to have a methodology that is able to split human and natural activities affecting air pollution due to the high seasonality of pollutant levels.

Studies that analyse the air quality and tourism indicators find that these indicators are closely related [50, 51]. Using a vector autoregressive (VAR) model a study focusing on tourism demand and air pollution causality found that tourism growth could have a negative effect on air quality [39] in terms of the particulate matter in the air that were significantly increased by tourism demand growth.

2.2 Sustainable tourism in different contexts

There are different approaches to considering a tourism strategy to be sustainable. Every approach has its own variables and focus points that determine if tourism development in a region is sustainable. Understanding if a tourist area is sustainable is often done by looking at satellite-derived datasets captured over time to understand the spatial-temporal variations of these ecological variables in tourism-related areas. By understanding the causes around these variables through a data driven approach, the analysis of air quality prospects in this study could give researchers and practitioners in environmental domains the insights to better understand the challenges around sustainable tourism development.

Understanding sustainability of an industry sector involves a great understanding of how it uses its resources to provide its output. Measuring the decline and regeneration of these resources can give insight if a process in its current form can continue to operate without exhausting the inputs in depends upon. This view on sustainability of an industry only provides a partial picture of sustainability and assumes a clear relation between input and output without looking at the impacts on other variables in the environment it operates. The current view around sustainability is the one that is often referred to as ‘sustainable development’ including a wider range of variables related to environmental, economic and social aspects [5]. In the case of environmental analysis, understanding land use and land cover and the relationship with tourism analysis before the advent of Geographical Information Systems (GIS) and remote sensing capabilities, required field studies which often consisted of ground-bases surveys and census data [46]. In the context of air quality, using GIS tools like GEE could play a valuable role for the analysis of the effect of tourism demand on air quality. A study that analyzes human activities on landscapes for instance, used GEE to understand the cumulative impacts of these activities on various scales [13]. The researchers indicate that due to the sparsity of such cumulative effects, technologies around earth ob-

servation and computation techniques are needed for assessing the human activities, in this case those around tourism, and the influence following this modification on climactic and environmental characteristics [13].

The chosen aspects that are researched in literature also determine the use of their tools like GIS. When focusing on recreational potential of tourist destinations in nature areas, it is found that GIS software can be used to measure the likelihood of area visits based on distance measures of attractions to better understand the risk for potentially damaging vulnerable landscape environments by nature-based tourism pursuits [12]. This form of spatial analysis focuses mainly on a specific area where some data sources like local individual preferences for walking routes and other similar manually collected data sources through surveys are used. This shows that depending on the use case, GIS tools can provide the means of effectively analysing spatial data in combination with other sources of data on different scales.

A different form of sustainability looks at transportation of people in a tourist area. Here, accessibility to infrastructure facilities and their link to tourist attractions is the main point of interest. This form is of special interest for city planners, which could have the aim of optimizing the city layout to support ease of traveling. The tools that are used for this kind of sustainability analysis are usually specialised in route planning [38]. As effective transportation is tightly linked with the topology of cities and the placement of transportation facilities, software systems that make use of urban modelling techniques are important to understand complex city challenges around environmental sustainability and maintainability [18, 21]. Analysis of mobility patterns between tourism destinations and the place where tourist stay showed that tourists tend stay near the destinations they want to visit the next day [37].

2.3 Measuring methods for understanding tourism

Sustainability is a broad subject that involves great understanding of the environment in which it is measured. In the case of sustainable tourism, this research work studies the effect of tourism demand on air quality. In order to effectively understand the tourism demand on air quality, literature is broadly studied on the topic of tourism impact on the environment.

2.3.1 Measuring methods for environmental impact of tourism on air quality

Tourism can impact the environment through air pollution in different ways depending on the intensity. The affect of tourism on environmental factors is a topic that is widely studied with spatial satellite data looking at different environmental stressors. The land use of a tourism attraction takes in space which could have been a natural place before. In literature, this effect of transformation of the natural environment to non-natural space or the loss of natural space in general caused by tourism is often captured by looking at changes in vegetation levels [14, 42] emphasising the benefit of using technologies like GIS and remote sensing [11]. Besides the spatial configuration that tourism influences, tourism activity itself can affect other environmental factors like temperature and air quality. For temperature, there are different ways to analyse its change in a region over time. When looking at changes

in urban growth patterns caused by tourism, one can analyse the land surface temperature in tourism areas to understand the temporal localities for tourism activity on a seasonal basis [54]. Unlike temperature analysis, air quality is defined by different pollutants that are considered in the analysis. As the tourism industry is driven by other industries for its supply of energy, food and consumer products, it affects the different quantities of pollutants in the air and therefore overall air quality [48]. In order to effectively understand the effects of tourism on the environment, recent literature suggests that tourism should be seen as sub-sectors to better understand the impact of different activities within tourism itself like accommodation, transportation and entertainment [25].

2.3.2 Tourism indicators to be used for measuring methods of air quality

A common indicator of measuring the impact of tourism on the economy is by looking at the added gross domestic product per capita by tourism [26, 17]. Oftentimes the economic indicator and tourism are positively linked to each other, however the magnitude of this observation varies and is determined by the methodology used in the empirical study [10]. In general, tourism and travel can greatly contribute to the income and wealth of the residents by the presence of traveling people [20]. The activities that are performed within the tourism sector are in the end aimed at financial profit and understanding the relationship between tourism activity and sustainable economic return is valuable in order to maintain this economic prospect. Tourism as an industry can be complex in the sense there are many different stakeholders involved in performing activities related to tourism. Relationships between tourism related private and the public sectors are seen as valuable in the industry [35] and could provide substantial competitive benefit to a tourism industry [23]. In the light of sustainable tourism these relations between different actors does provide a challenge as in order to see effective steps to sustainable tourism all actors should be involved [22] which becomes challenging if the sustainable measures taken cannot maintain economic viability in the long-term [24].

Following this line of though, it is found that economic sustainability of tourism is dependant on the number of tourist arrivals [9]. As economic output of an industry like tourism is found to be linked with air quality levels [51], using tourism arrivals as an indicator for sustainable tourism could therefore be a promising feature for modelling the effect on air quality. The affect of an industry on the environment can be analysed through different methods. Significance testing is a method for statistical analysis of trends in the relation of air quality data at identified tourism sites. Correlation between air quality pollutants and tourism demand indicators could be analysed through Pearson correlation analysis [13]. In the context of the tourism industry, regression analysis can be used to understand the relationship between tourism indicators and air pollution [40] more effectively.

2.4 Implications for the current study

From the literature that was analysed several interesting technologies and methods showed their potential in analysing the effects of tourism demand on air quality. For example these are GEE and its capability of collecting data on a large scale. In this context of analysing

the air quality in Italy based on tourism demand, this is beneficial as it is able to provide data on the entire country. At the same time, it contains widely available data which is also continuously updated, providing researchers and practitioners the ability to conduct similar studies. It was also found that the use of tourism indicators like number of arrivals can be used to effectively model tourism demand in the context of air quality analysis. Combining this with the ability to use RDF technology to systematically retrieve tourist locations using SPARQL queries is a useful tool with a multitude of research applications. Specifically, in this study this allows to collect locations of different types of tourist sites which are important to broadly model the impact of the tourism industry.

Chapter 3

Methodology

This chapter discusses the study area along with the data sources for analysing the air quality around tourism sites in Italy. The retrieval process of the data as well as the pre-processing required for the machine learning analysis is also discussed. The construction of the data needed for this machine learning model through a data processing pipeline is also addressed.

3.1 Study area and data sources

This section defines the study area of this work together with the data sources that are used for the air quality modelling. The description of the data is provided together with the sample frequency of the datasets.

3.1.1 Study Area

The area of study in this work is Italy which is located in Southern Europe in the middle of the Mediterranean Sea. It has a population of almost 59 million people in 2022 [34] and an area of approximately 301,230 km^2. Italy lies within latitudes 35.2890 and 47.0921 and longitudes 6.6273 and 18.7845. Tourism is an important part of the Italian economy as it contributes to 9.1 % of its GDP in 2021 [43]. In 2019, before the impact of the global pandemic, Italy received around 95 million tourists [6].

In the context of air quality analysis, the air pollution in Italy differs across different parts of the country with northern regions being more polluted compared to other regions [2]. When performing air quality modelling using tourism demand, the areas where there are other significant anthropogenic factors or geographical circumstances that affect air quality distribution need to be carefully assessed by considering the possible impact of these other sources of pollutants. Air quality modelling on a regional level with varying levels of external sources of air pollution could be used to compare the patterns that are found in highly polluted regions with regions that are less polluted. This way the impact of tourism demand on air quality can be understood from a measurability standpoint as well in the presence of confounding sources of air pollution.

3.1.2 Data Sources

The analysis performed in this study is based around several datasets formed by accessing ISTAT, GEE and the Wikidata knowledge base. The data from these datasets are used to train machine learning models to model the link between tourism arrival rates, nights spent and the average length of stay on the air quality levels at different categories of spatial locations. These spatial locations represent the places where tourism traffic use is expected to have an influence on the air quality. The metrics used for air quality measurements are therefore based on the types of pollutants caused by transportation use. Being able to model the effect of changing tourism demand both on a national and regional level can provide valuable insights to evaluate the sustainability practices of different regions.

Air quality analysis is performed by using CAMS and Sentinel-5P products from GEE. The CAMS product has the advantage that the data is in situ [36], meaning that the data is formed of a combination of different data sources like ground level measurements together with satellite data to obtain better calibrated and verified measurements. The CAMS and Sentinel-5P products are chosen based on their relative modern equipment and their recent availability of data which connects well with the availability of global tourism numbers per region from ISTAT which they started publishing a since 2016. The regional modelling using both satellite products for air quality sampling is performed to better understand the relation between tourism demand indicators from ISTAT and air quality levels.

The analysis of air quality levels is based on the types of pollutants contributed by transportation and is conducted at different types of locations related to tourism. Particulate matter less than 2.5 μm ($PM_{2.5}$), carbon monoxide (CO), nitrogen dioxide (NO_2), sulphur dioxide (SO_2) are used to assess the air quality levels based on their frequent links with transportation in literature. The sites at which these pollutants are measured can be placed in three different categories which are: accommodation sites, transportation sites and tourism attractions which are retrieved from the Wikidata knowledge base. In Table 3.1, the primary data sources used in this research for modelling the effects of tourism demand on air quality are listed.

3.2 Data retrieval and pre-processing for feature engineering

In order to effectively train the machine learning models it is important that the required data sources for the learning process are in a suitable format. This work makes use of different datasets spanning different time periods to construct a training set and for this a clear strategy for pre-processing and combing the different data is needed. This section goes into the methods used for the retrieval, most notably making use of the RDF data model for the retrieval of categorical tourism sites, and the preparation of the data by explaining the final format which is used for the learning process.

3.2.1 Retrieval and pre-processing of tourism demand data from ISTAT

A central data source of this study is ISTAT which gives insights in the monthly tourism statistics in Italy. The goal is to retrieve tourism demand data which is valuable for forming

Content type	Data source	Data type	Resolution / Level	Time span
Copernicus Atmosphere Monitoring Service (CAMS)	Google Earth Engine	Raster/Satellite imagery	44528 meters	07.2016 12.2021
Sentinel-5P Satellite	Google Earth Engine	Raster/Satellite imagery	0.01 arc degrees	07.2018 12.2021
Tourist arrivals, nights spent and average length of stay	Italian National Institute of Statistics	Numerical data	National and regional level	07.2016 12.2021
Tourism points of interest data	The Wikidata Knowledge Base	Spatial objects	National and regional level	-

Table 3.1: Data sources used for modelling air quality impact based on tourism demand.

a link with the air quality data from the GEE platform which is discussed later in this section. The amount and granularity at which the data is available is an important factor for the performance of the model as well as the features on which the learning takes place.

From the data portal of ISTAT depicted in Figure 3.1, it can be seen that the statistics institute offers a number of indicators about tourism which could be used as features for the machine learning modelling. The statistics are categorised under 'Capacity of collective accommodation - municipality data', 'Occupancy in collective tourist accommodation - monthly data', 'Occupancy in collective tourist accommodation - yearly data', 'Gross and Net rate of bed-places in hotels' and 'Tourism indicators - monthly data'. As the aim of the modelling process in this research is based on tourism demand, numerical data on the number of tourist arrivals, nights spent and average length of stay are the main data values that are used. This was chosen as the capacity indicators of accommodation sites indicate less about the direct inflow of tourists and their effect on air quality in general. The feature vectors that are used for the machine learning modelling are therefore arrivals, nights spent and average length of stay by the tourists. As the CAMS dataset is available since July of 2016 in GEE, the tourism data is used from that date on.

When retrieving the tourism demand data on different administrative levels, challenges can arise that require additional pre-processing to effectively use the data for the analysis. This was the case for the Trentino South-Tyrol region where the regional data on tourism arrivals and nights spent was partly split for Trentino and South-Tyrol. This data had to be combined manually to be used as one first level administrative region for the analysis in this study. This case will be further discussed in the following chapter in a general setting.

3.2.2 Retrieval of study locations for air quality modelling from Wikidata

The location data used for this study to identify relevant tourist sites is retrieved from the Wikidata platform which is a collaboratively made knowledge base [45]. The locations of

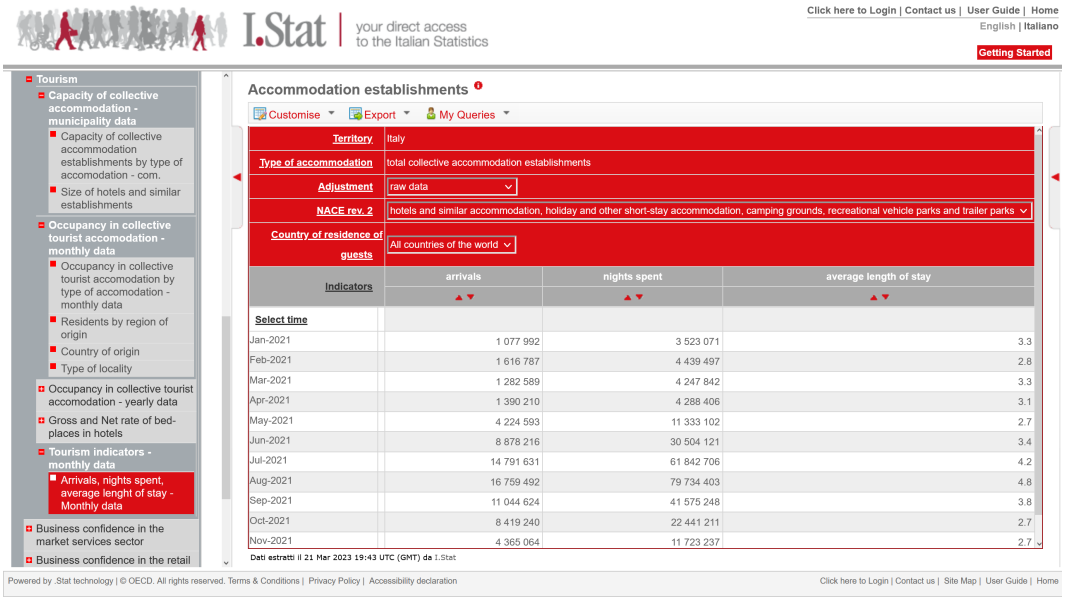

Figure 3.1: The ISTAT data portal with data about the tourism industry in Italy.

interest are retrieved from the knowledge base by constructing a SPARQL query using the entities and properties resembling the search for tourism related sites in the accommodation, transportation and tourism attractions categories. These three categories are chosen based on their key role in regular transportation patterns of tourist between their residence place and the destinations they want to visit through the use of public transport nodes.

Research that has used mobile phone data to analyse mobility patterns between tourism destinations, the place of stay and transportation hubs indicates that tourists tend stay near the destinations they want to visit the next day [37]. The study conducted in a city environment found that central transportation hubs are highly popular as a means of transportation and that the tourist destinations are spatially correlated with these public transit facilities. Similar research on mobility patterns indicates that tourists want to visit many amenities during the length of their stay within a reasonable amount of time [44]. Findings involving the analysis of mobility patterns of tourist groups also suggest a high use of transportation to reach the places the tourists want to visit [49]. To broadly understand the effect of tourism on different types of tourism locations, the following sections go into the process of querying the locations of the accommodation, transportation and tourism sites for performing the air quality sampling. In Figure 3.2 the process of visualising the queried locations can be seen using the build-in map of the Wikidata Query Service.

Accommodation locations retrieval

Based on the data available from the ISTAT database which uses the number of people arriving at hotels and other tourist accommodation places, the SPARQL queries contain the entities which represent the types of accommodation used by the statistics institute. Per-

Figure 3.2: The retrieval process of study locations in the Wikidata Query Service showing categorised locations.

forming the search on hotels and campsites in Italy produces an initial 3049 results. In the process of finding additional results for hotels, the accommodation facilities in Wikidata that have a Booking.com identifier attached to it were queried as well. After performing an union operation on the results retrieved from the hotel and camping site query, the results increased with another 12 results to a total of 3061 results after excluding a match located outside of Italy. The resulting entities are ones that are listed in Wikidata with a Booking.com ID without being a direct instance of a hotel. When these additional results were analysed separately by performing a subtraction operation with the original set of hotels, it was found that these results are of types like vacation home, hostel and guest house. As these different types of entities are found to be all of the type hotel, the predicate of the query was adapted to include both objects that are an 'instance of' (P31) a hotel or are at any level in the subclass tree of a hotel by making use of the 'subclass of' property specified

as P279*. This resulted in a substantial amount of more results of 3714 entities covering different types of accommodation facilities which fits well with the description in the ISTAT data portal of the type of accommodation that is used to count the number of arrivals which specifically mentions 'similar accommodation' to hotels. Besides these accommodations that are similar to hotels, the description of ISTAT also mentions camping grounds, recreational vehicle parks and trailer parks. When entities were used to also retrieve these locations, only queries for campsite gave additional 11 matches.

Important to note in this process is that many locations in these accommodation categories are not tagged with information that could be used to retrieve the Italian region in which they are located which is needed for the regional modelling in this study. Of the 3725 results obtained using the described method, only 557 results had a match for their 'located in the administrative territorial entity' property (P131). Furthermore, the regions that were retrieved were at various administrative levels like municipal and commune levels instead of the first-level administrative divisions of Italy. There exists a property Wikidata data resembles Italian municipal regions through the special ISTAT ID (P635). However, this identifier is rarely used producing significantly lower results compared to matching the more common 'located in the administrative territorial entity' property.

For this research the region plays an important role as the machine learning analysis of the effect of tourism demand on air quality is performed on both a national and regional level. To overcome this challenge, the exported data without querying the regions was used together with the GeoPy Python package to add the regions to each entity by making use of the entity coordinates retrieved from Wikidata.

Transportation locations retrieval

Retrieving the transportation locations was similar to the process of retrieving the accommodation places. The query still consisted of querying Italy by retrieving the coordinates of each entity. The objects queried for are related to transportation facilities that are found to be important for the tourist sector. Railway stations and bus stations were queried as well, as they were seen as important sources of transportation to visit different amenities during a tourist trip. For querying the airports an initial challenge occurred that there were airports in the results that are not used for civil aviation. To overcome this challenge, the query was adapted to retrieve only the airports that have an International Civil Aviation Organization (ICAO) code attached to it.

Tourism attraction locations retrieval

The approach for retrieving tourist attractions is more subtle than retrieving locations from the other two categories as many things could be considered as tourist attractions. The object identifier for tourist attraction (Q570116) gave 68 results when also querying for the region which would be a limited sample for analysing tourist attractions in Italy. In order to still have a greater set of tourism attractions that represent places that tourists are likely to visit frequently, a similar approach was performed to using the Booking.com ID as a property in the query. This availability of external identifiers representing entities in

Category of tourism site	**Description**	**Wikidata object identifiers and properties**	**Number of entities**
Accommodation facilities	Tourist destinations where people can stay overnight.	Hotel (Q27686) Campsite (Q832778) Booking.com ID (P10532)	3725
Transportation locations	Transportation locations often used by the tourism sector.	Airport (Q1248784) Railway station (Q55488) Bus station (Q494829)	3554
Tourism attractions	Tourism attractions that people can visit.	Tourist attraction (Q570116) TripAdvisor ID (P3134) (Excluding ones appearing in previous categories)	1327

Table 3.2: Wikidata categorised tourism locations.

different systems is a valuable characteristic of the semantic web that enables the linkage of different knowledge graphs. Similar to the Booking.com IDs there also exists a reference to the TripAdvisor IDs in Wikidata. The use of this property in the query thus allows to retrieve all locations in the Wikidata knowledge base that people have attached a TripAdvisor ID to. The approach of using the TripAdvisor ID gave a total number of 1286 results when also querying for the regions. Unlike the case for retrieving the regions of the accommodation places, the results for the entities with TripAdvisor IDs attached did have region information attached to the vast majority of the results. However, it was still the case that the resulting regions were on different administrative levels. Although there exists a Wikidata item for the first-level administrative regions of Italy, namely Q16110, some of the regions are not tagged with this item. To overcome this, a 'values' statement is used containing all 20 first-level administrative values in Wikidata which was combined with the 'located in the administrative territorial entity' property specified as P131* to match one of the regions in the values statement. In this approach, the query finds the region by going up the tree of administrative divisions to eventually find the first-level administrative region which is then combined with the output for each entity.

Retrieving locations in this category as well as the two others resulted in duplicate locations for some of the locations. Upon further inspection of these duplicate entries, it was found that they had different set of coordinates attached which are slightly different from each other. As these locations are generally static entities in space, a possible explanation could be that they were added to denote different buildings of the same entity which explains the close proximity of the multiple coordinates. To address this, the queries for retrieving the tourism locations were adapted to sample the set of coordinates in the select statement and to group the remaining columns using the 'group by' statement.

The final overview with the number of locations retrieved per category is found in Table 3.2. A short description and the use of entities and properties important for tourism attractions are also provided in the table.

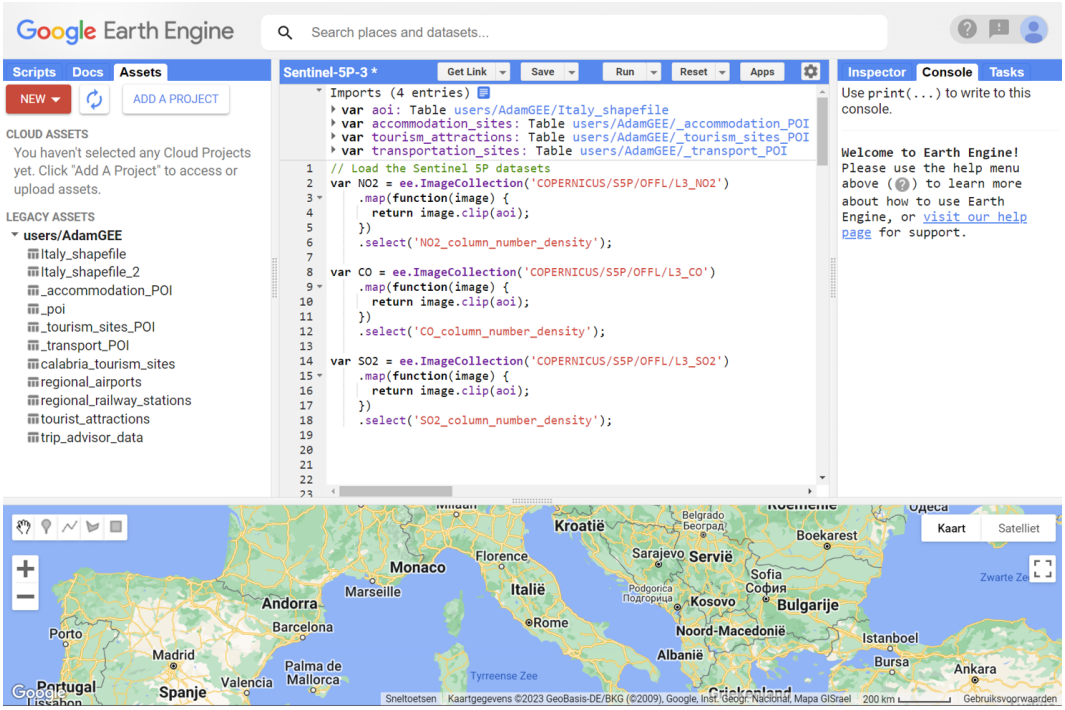

Figure 3.3: The GEE dashboard showing Sentinel-5P data being imported for analysis of Italy.

3.2.3 Air quality sampling at identified locations using Google Earth Engine

Having retrieved the data on tourism demand, the air quality levels at the tourism locations form final target variables which the machine learning modelling aims to predict. The exported data from Wikidata contain latitude and longitude values which are provided for each entity retrieved by the query. This data file, which is a CSV file, is imported to the GEE platform for the analysis of the locations after specifying the correct columns in the file for the latitude and longitude. The GEE platform contains a code editor which is used to specify how the datasets from different satellites are imported and analysed by sampling the locations which are imported by the user. For example, in Figure 3.3 the GEE dashboard shows Sentinel-5P data being processed.

As described in the previous sections, two GEE products are used for retrieving the monthly time series air quality values. These are the CAMS and Sentinel-5P products which each have their own relevant bands for modelling different pollutants. The bands that are selected are related to the pollutants caused by the use of transportation as these are associated to be a major contributor of air pollution found in literature. The use of transport is therefore the primary effect of tourism on air quality that this work models. For the CAMS product, the band for the $PM_{2.5}$ was most useful for this analysis based on its frequent use and relation with transportation analysis.

The Sentinel-5P is in several ways different from the CAMS satellite, both in its spatial

GEE product	Pollutant	Description	Satellite product band name
CAMS	$PM_{2.5}$	Particulate matter with diameter less than 2.5 µm	particulate_matter_d_less_than _25_um_surface
Sentinel-5P	CO	Carbon Monoxide	CO_column_number_density
	NO_2	Nitrogen Dioxide	NO2_column_number_density
	SO_2	Sulphur Dioxide	SO2_column_number_density

Table 3.3: The bands used for the CAMS and Sentinel-5P products in GEE.

resolution as well as the pollutants it can measure. Equipped with a spectrometer called Tropomi (Tropospheric Monitoring Instrument) Sentinel-5P can measure important pollutants for air quality analysis including nitrogen dioxide and sulphur dioxide which are caused by transportation as well as other anthropogenic factors. The pollutants that are chosen are based on their link with combustion of fossil fuels to capture transportation use by the tourism industry. For each GEE product used, the bands are listed in table Table 3.3.

The analysis of air quality is performed on each category separately to better understand the differences of air quality levels among these categories. This could capture the different affects of the tourism industry on air quality in line with the related work. These related studies indicated the effect of vehicles on air quality, and the sampling of air quality at the transportation and accommodation sites in this study could therefore provide insights on the effect of tourism demand.

3.3 Data integration overview with data pipeline workflow

In Figure 3.4 an overview is given of the data sources used together with the research methods applied. Using different sources of data for the task of machine learning modelling requires that the data is in the right format to enable the learning process. This is achieved through creating different stages of capturing and processing of the data which eventually leads to a data pipeline where the final datasets that are created are analysed. This work uses an LSTM model from the Keras library to perform air quality prediction based on the data from the data processing pipeline. As multiple pollutants are part of the analysis, the data from the different satellite products is joined with the monthly tourism statistics. The way these integrated data sources are stored in data structures to perform effective analysis is important. The different dimensions of data that are considered should be stored in such a way that the modelling and visualisation process of the analysis benefits from the way the integrated datasets are stored. The Pandas library is used to capture the data in dataframes making it fit to be used conveniently with the Keras library. The national level dataframe contains the columns for the tourism arrivals, nights spent and average length of stay data together with the air pollutant levels for each of the three different tourism categories. The regional level dataframes use similar structuring while keeping track of which region the dataframe is about in code of the modelling. The Matplotlib visualization library is used to

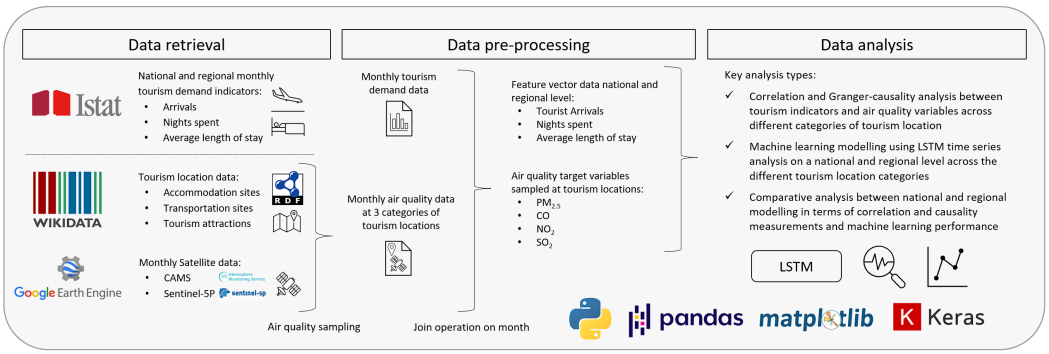

Figure 3.4: Research workflow showing data sources and dataset creation pipeline.

visualize the relationship of tourism demand features and different air pollutant levels to get a better intuition for how the data behaves. LSTM air pollution prediction is visualized as well after the learning process using the same visualization library.

Understanding how national patterns of the effects of tourism on air quality relate to the findings on a regional level under different external air quality conditions is an important part of this research. This would give better understanding of how the tendencies of the effect of tourism on air quality are shaped across regions. This would show to what extent the national level analysis can represent regional based measurements. The data processing pipeline is utilized in the same way as for the national analysis through filtering the datasets on the studied regions first. Only the categorical results of the Wikidata elements of a specific region are then considered and sampled with satellite data. These air quality levels and the regional tourism demand data from ISTAT are then turned into a single dataframe by performing a join operation on the date columns of the two datasets. Considering the different start dates of the satellite products used for the air quality sampling, this join operation therefore creates the dataframes that conveniently match the availability of the air quality data with the tourism indicators from ISTAT. For the three categories of tourism sites there are dataframes made that hold the different pollutant levels at the respective locations the categories represent. Thus, the output of the data pipeline are datasets on which machine learning models can learn the relationship between tourism demand indicators, such as tourist arrivals and nights spent, and air pollution data on a national and regional level across the different categorical tourism sites. These datasets are also used for the correlation and causality analysis between the tourism features and air pollutants which is discussed in the next section.

3.4 Correlation and Granger-causality analysis

The data integration process described by the data processing pipeline produces the datasets that are used in this study to analyse the effect of tourism demand on air quality. The first type of analysis that is performed on the resulting datasets is the correlation analysis

between tourist arrivals and tourist nights spent of the considered pollutants. Performing this analysis on national level data gives an initial view on how the arrivals and nights spent time series relate to the pollutant time series. Based on this information, this study continues the analysis by focusing on the pollutant time series which are found to be most correlated with the tourism demand indicators. This is done by eventually modelling the relationship using a machine learning model on a national and regional level and comparing the performance between these levels and across different regions.

Correlation between the tourism demand indicators and the pollutants only give one side of the story when analysing the effects of tourism demand on air quality. It is important that one can verify that the tourism demand time series provides significant information to predict the pollutant levels that are analysed. The Granger-causality test provides a more stringent measure to analyse if one time series can be used to model another and is often used for analysing the effect of tourism demand on different time series [7, 39, 28]. After conducting a correlation analysis to determine if a pollutant is correlated with tourism demand, the Granger-causality tests is used to further analyse the relation between the tourism demand indicators and the pollutant time series across the different categories of tourism sites that are analysed. This gives an indication on how well it can be used for the LSTM modelling process given the results for the different lag times. The p-value resulting from the analysis between each tourism demand indicator and the pollutant indicates if including the tourism demand information at a given lag value can provide significant information to model the pollutant.

The regional differences in terms of correlation and Granger-causality values can give information on how well regions can be modelled in different parts of the country. Combining this information with the knowledge about average distribution of air pollution in the country can give insight on the effect of different air quality levels on the modelling performance of tourism on air quality. For this analysis, the feature collection objects that contain the tourism locations in the three categories are filtered on their region property. Three assets are formed in the GEE platform by loading in the resulting files representing the three categories. By using the asset in a feature collection object, the collection can be filtered on the region that is sampled by passing a GEE filter object to the feature collection containing the particular region name. Three GEE list objects are created containing the regional sampled air quality data for each of the pollutants to be analysed which are subsequently exported for the use as target variables for the machine learning analysis.

3.5 LSTM modelling of tourism demand and air quality time series

At this point the national and regional data needs to be transformed to feature vectors such that the LSTM model can use these to learn patterns in the data and to see if the correlations measured are reflecting the modelling performance. The features that are used for the learning process are based on the tourism indicators retrieved from the ISTAT database which are tourism arrivals, nights spent and average length of stay. Besides these features, the year and the month are also added as separate features. The motivation behind this

choice comes from the fact that seasonality plays an important role in the tourism industry. Using monthly time series of tourism demand data could therefore allow the model to learn seasonal trends to incorporate the increases during the holiday seasons. Based on this non-linear and stochastic nature of tourism flow [29], this research is using the LSTM model to capture and learn from these non-linear trends in its input data. The target value of the modelling will be the air pollutant levels across the three categories of tourism locations that are sampled. The input features are normalized for better model stability due to the varying feature sizes.

The machine learning modelling process in this work is categorized by the two types of spatial granularity following. First, the correlation between the tourism demand indicators and the national level data on air pollutants are calculated. This calculation is performed on the three tourism categories separately to better understand the effect of the tourism industry on the different categorical tourism sites. Second, the regional level analysis follows the same approach as the analysis on the national level. Italy contains 20 first-level administrative regions and the same tourism indicators for the national level are available for the regional level. The benefit of the regional analysis over the national analysis is that it can better capture local patterns due to the data covering a smaller spatial area. Performing regional analysis allows to better reduce the noise from other factors which can affect air quality as well such as other anthropogenic factors like industry activity. Analysing the relationship between tourism and air quality across different regions with different air quality patterns can also give insight in how strongly the effect of tourism on air quality is measurable in different environments.

3.6 Evaluation approach of national and regional LSTM analysis

The machine learning modelling will be evaluated both on a national and regional level where the regional analysis results in terms of predictability across regions are compared with each other to understand the modelling performance under different air quality conditions. The LSTM analysis is based on the tourism categories and pollutants that are found to be correlated the most. Evaluation of the performance of the trained models is done by using the root mean squared error (RMSE). This evaluation metric is chosen based on its clear interpretation for prediction analysis given its purpose to measure the square root of the average difference between the actual and predicted values. A training, validation and test set will be constructed which are used to train and assess the LSTM model. After training the model it will be evaluated in terms of how well it can predict the air pollutant time series based on the three tourism demand indicators of arrivals, nights spent and average length of stay on a national and regional level.

Chapter 4

Results

In this results chapter, the data that is collected through the methods described in the methodology is used for the modelling process. As described, the goal is to perform analysis to see how well air pollution can be modelled based on tourism demand data. The construction of machine learning models in this context as well as their performance analysis are further topics that are discussed in this chapter.

4.1 Data integration process and the generalized challenges

This section discusses the challenges that arose when performing data integration. It was chosen to discuss the challenges that could be generalized to occur in similar data analysis studies which could be valuable to take into account when performing data integration.

4.1.1 Variability and generalization of temporal data in heterogeneous data integration

The data integration that was described in the methodology allows the creation of the datasets needed in order to perform the analysis between tourism demand and air quality. This integration process is described in the data processing pipeline that was shown in Figure 3.4. As this integration happens over heterogeneous datasets that differ based on spatial precision and temporal scale, these datasets need to be aggregated in order to be useful in the analysis. The difference in temporal scale requires that data from one source that is on a smaller temporal scale is adapted to work with the data on a larger time scale. In this work, this was the case for the data retrieved from the CAMS and Sentinel-5P products which provided data on a fine temporal scale. As the data provided by ISTAT is monthly tourism data, the measurements of the two satellite products were aggregated to this monthly time frame. For the CAMS product, the $PM_{2.5}$ data is provided on an hourly time scale. The Sentinel-5P satellite product provides its data on a daily time frame. This means that some of the expressiveness and variability gets reduced in the aggregated monthly data, making it more challenging to derive fine-grained information like weekly trends without using additional assumptions about the tourist flow per month.

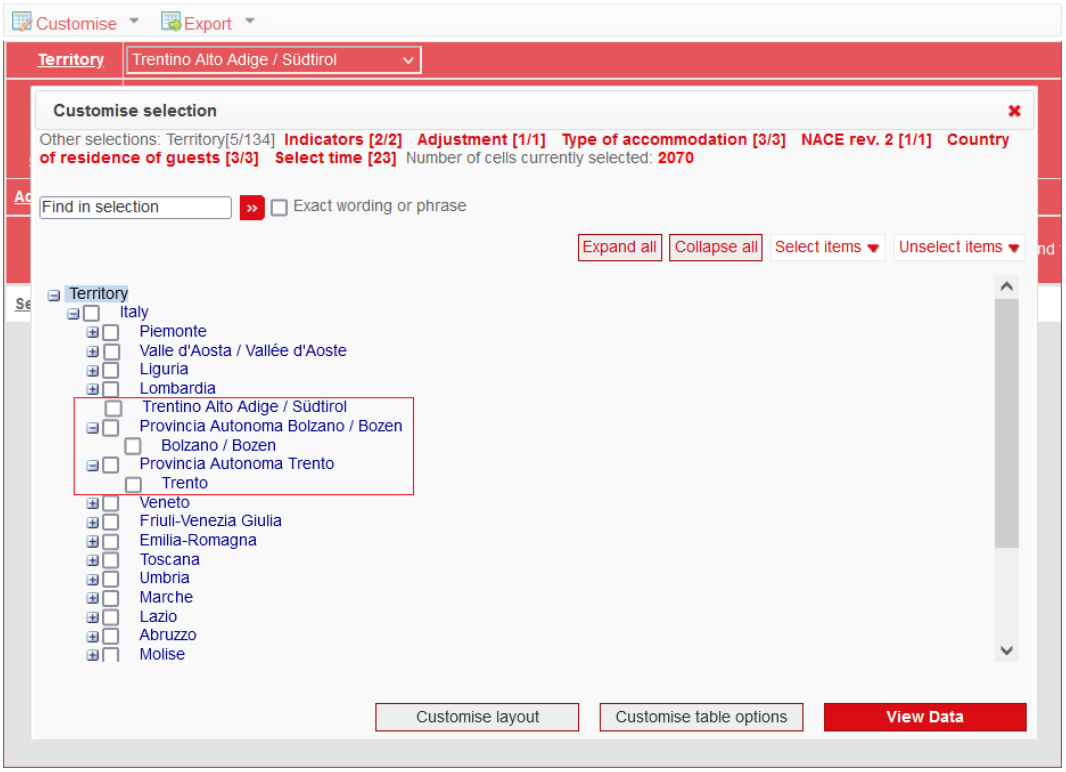

Figure 4.1: Retrieval of first level administrative division data from ISTAT showing Trentino-Alto Adige/Südtirol also as Provincia Autonoma Bolzano and Provincia Autonoma Trento.

On the other hand, as tourism data is highly seasonal the aggregation in this work leads to increased generalization. It allows the analysis to derive general trends across the different categories of tourism locations without focusing too much on the smaller temporal patterns. It allows to have a more focused scope on the overall patterns considering the multidimensional nature of the analysis being performed across both national and regional scale and the different categories of locations.

4.1.2 Data pre-processing of administrative divisions

The retrieval of the tourism demand features that this work uses is on a national and regional level. For the regional level analysis, data on first-level administrative divisions are available on the ISTAT portal. As Italy has 20 first-level administrative regions, it was expected that this number of regional entries where available for retrieving the tourist demand data from. However, when having to selects the regions it was found that 22 regions were given inside the portal. Upon further observation it was found that the first-level administrative region Trentino-Alto Adige/Südtirol was also given separately as Provincia Autonoma Bolzano and Provincia Autonoma Trento which is shown in Figure 4.1.

During the export of the tourism demand data, entries were given for these three territories separately. Data on Trentino-Alto Adige/Südtirol was only provided partly with data going back more years being split across the data entries for Provincia Autonoma Bolzano and Provincia Autonoma Trento. Based on the data that was provided for the combined level of Trentino-Alto Adige/Südtirol, it was found that the arrivals and nights spent data could be added from the two subdivisions to have values for the top-level Trentino-Alto Adige/Südtirol region to use further in the analysis.

Such an occurrence could also be found in other databases that provide data about administrative divisions. This depends on how administrative divisions are structured and how they provide their statistics about the area they govern. Potentially deviating methods on how they collect and present their data could introduce more steps in retrieval process and could lead to additional pre-processing practices as a result. Knowing how to perform the manual pre-processing often requires additional knowledge about the situation at hand like in the case of Trentino-Alto Adige/Südtirol. In this case, knowledge about the region having a special status of autonomy and that their two constituent provinces provide their data separately helped to solve the given case.

4.2 National level data exploration through visualization and correlation analysis

After collecting the data, monthly data about tourist arrivals and nights spent can be loaded into a dataframe on both a national and regional level. The same holds for the different pollutants that are measured at the different categories of tourism locations. Following the approach given in the methodology, these tourism features and pollutant levels are combined in one dataframe which allows for convenient visualization of the data.

4.2.1 Tourism demand and pollution levels data exploration through visualization

In a data analysis study, visualizing the data can give valuable clues on how different datasets are related to each other and can reveal valuable patterns. The scatter plot grids in Figure 4.2 and Figure 4.3 are made which indicate for each pollutant and tourism location category the trendline in the data with regards to the tourism demand indicator. The two demand indicators in this case are the tourism arrivals and nights spent, here the average length of stay was excluded from the visualization as the metric results form the other two demand indicators through division. It can be seen that the $PM_{2.5}$ values are positively correlated with the two tourism demand indicators. Looking at the regression line for this pollutant, for all the three location categories the same kind of positive trend can be seen. The slope of the regression line being statistically different from zero is used as the null hypothesis in the linear regression function of the SciPy library. The specifics of calculating the p-value from this approach is based on a linear regression model fitting the data points represented as a combination of the tourism demand indicator and the pollutant and subsequently deriving the p-value from the slope coefficient. For each scatter plot in the grid,

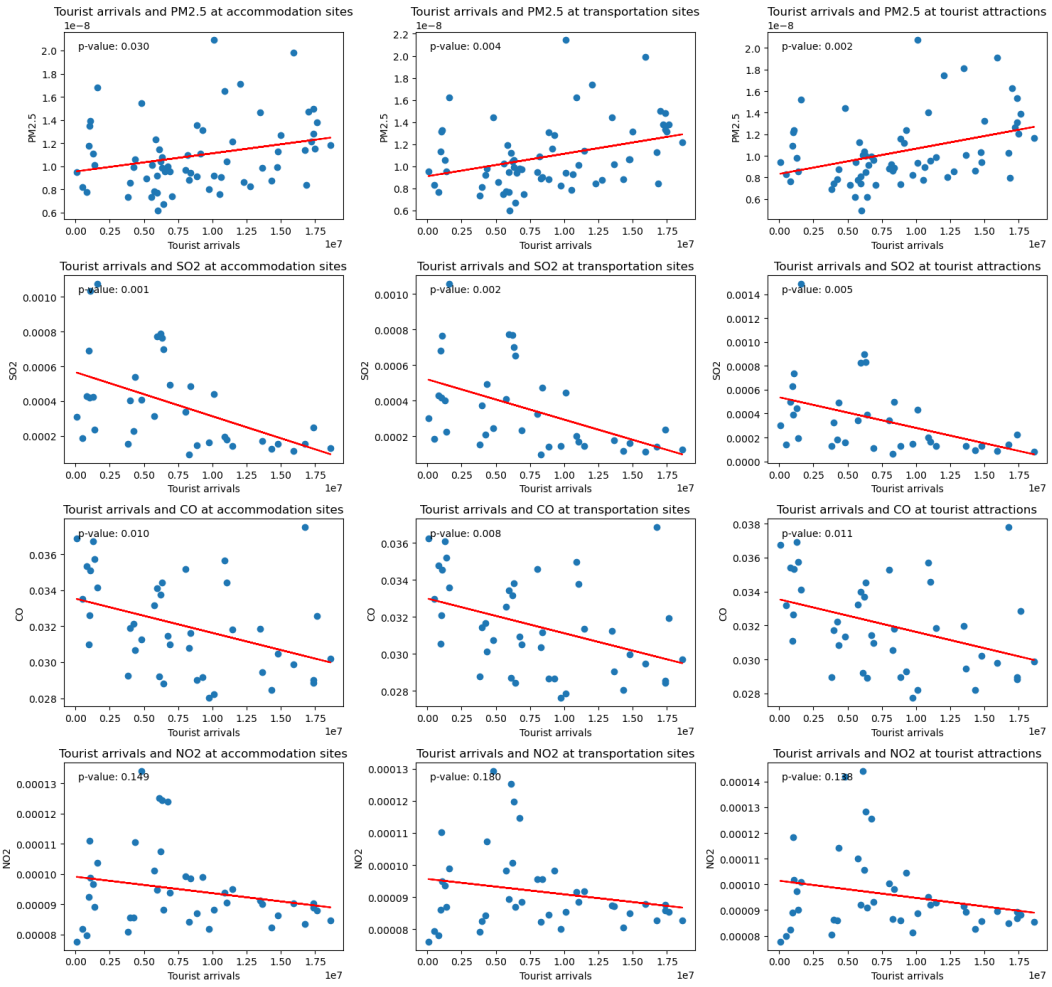

Figure 4.2: Scatter plot grid of pollutants and tourism arrivals on a national level for each tourism location category showing the trend line in the data and the p-value of the trend.

the p-value gives insights in the statistical significance of the relationship between tourism arrivals and the pollutant level in each of the three location categories. This relationship is denoted with a trendline for each of the pollutant and category combination.

Looking at Figure 4.2, the $PM_{2.5}$ pollutant captured through CAMS shows a positive trend for the three categories of tourism sites and looking at the p-value for the relation between tourist arrivals and the $PM_{2.5}$ levels at the three categories, it is likely that they are statistically related to each other. The three pollutants, SO_2, CO and NO_2 are found to show a downward trend in the data. For all the scatter plots in Figure 4.2, the trend is found to be significant either upwards or downwards except for the rends with NO_2 and arrivals in the bottom row of the figure. The downward trend in the data can be explained with the results in Figure 4.4 where of the pollutant from the two satellite products together with their minimum and maximum variation across the different categories of tourism sites

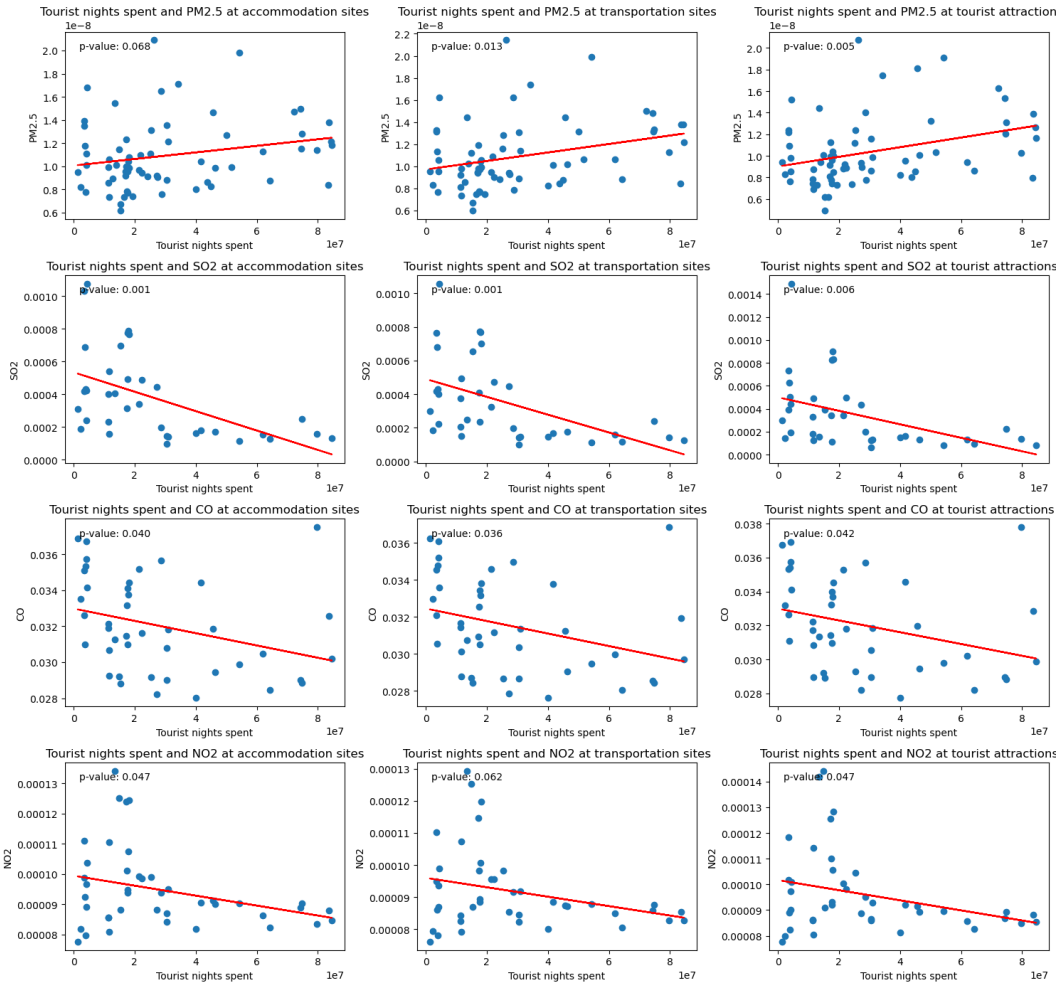

Figure 4.3: Scatter plot grid of pollutants and tourism nights spent on a national level for each tourism location category showing the trend line in the data and the p-value of the trend.

can be found together with the tourism demand indicators. From this figure, the categorical average of the pollutant levels is found to be higher at the times where arrivals are showing lower values.

Overall, the trends found with arrivals and the pollutant level are similar for the analysis performed with nights spent as the tourism demand indicators. In Figure 4.3, the trends in the relationship between nights spent and $PM_{2.5}$ pollutant are positive and compared to the analysis with the arrivals are a less significant. For the relationship between nights spent and the $PM_{2.5}$ pollutant sampled at the accommodation sites, the p-value is measured above 0.05, whereas the p-value in the case for the arrival's indicator did show a significant relationship. When looking at trend at the sampled transportation sites and tourist attractions it still shows a significant trend similar to the case of using the arrivals indicator. An interest-

ing observation can be made for the $PM_{2.5}$ pollutant where the p-values are the highest for the combination with the accommodation and the lowest for the tourist attraction sites with measurements of the transportation sites being in the middle of these two values. Although this could be in the margin of error, it does reflect the general tourism patterns in the sense that tourists are distributed among the different accommodation sites and travel to the transportation sites in order to reach the tourism attractions they want to visit. In this general view of tourist distribution and their mobility patterns, it could be explained that the tourist intensity is the highest at the tourism sites and therefore showing a stronger relationship. The transportation use of the tourists to come to the same place could be a potential cause of this stronger relationship of $PM_{2.5}$ at the tourism attraction sites, a similar cause could also be that the facilities around the tourism sites should be supplemented with resources through transport to serve the tourism industry like food related areas.

When comparing the results from the $PM_{2.5}$ values with the pollutants from the Sentinel-5P product, a different picture is seen. Here the pollutants are found to be negatively related to tourism demand which could have different explanations. The contrast found between the pollutants of the different satellite products could be caused by the in situ nature of the CAMS product for the $PM_{2.5}$. In situ data provides better calibrated and verified measurement for more accurate sampling of the pollutant at the three categories of tourism sites. Another reason could be that the significance of the tourism industry on Sentinel-5P pollutants are less than the effect shown on the $PM_{2.5}$ levels. The relationship between SO_2, CO and NO_2 and tourism demand could be more challenging to visualize due to other more significant factors like overall energy use in colder seasons explaining the higher peaks in Figure 4.4. The analysis is continued on a regional level for the $PM_{2.5}$ pollutant based on the higher correlation levels found on a national level. This could give more insight on how the national level relation between tourism demand and $PM_{2.5}$ relates to the results on a regional level. The correlation results and Granger-causality results will be compared for the accommodation, transportation and tourist attraction locations. How the increased correlation at the tourist attraction sites related to the other two location categories found in Table 4.1 will also be compared with the regional level.

4.2.2 Correlation analysis of tourism demand and air quality pollutants

This section looks at the different bands that were chosen in the methodology that could be impacted by tourism demand. In Table 4.1 the Pearson correlation of the pollutants sampled at the different tourism attractions can be seen. From these results, it can be seen that the $PM_{2.5}$ pollutant from the CAMS product has a positive correlation with tourism arrivals and nights spent. The correlation results for the Sentinel-5P pollutants are negative which are in line with the scatter plot figures for both tourist arrivals and nights spent and the line plots in Figure 4.4.

As described in the previous section, the analysis of the relationship between tourism demand and air quality is continued with $PM_{2.5}$ on a regional level. To understand how well the tourism demand indicators can be used to predict the $PM_{2.5}$ pollutant at the three categories of tourism locations, the Granger-causality test is conducted in section 4.4.

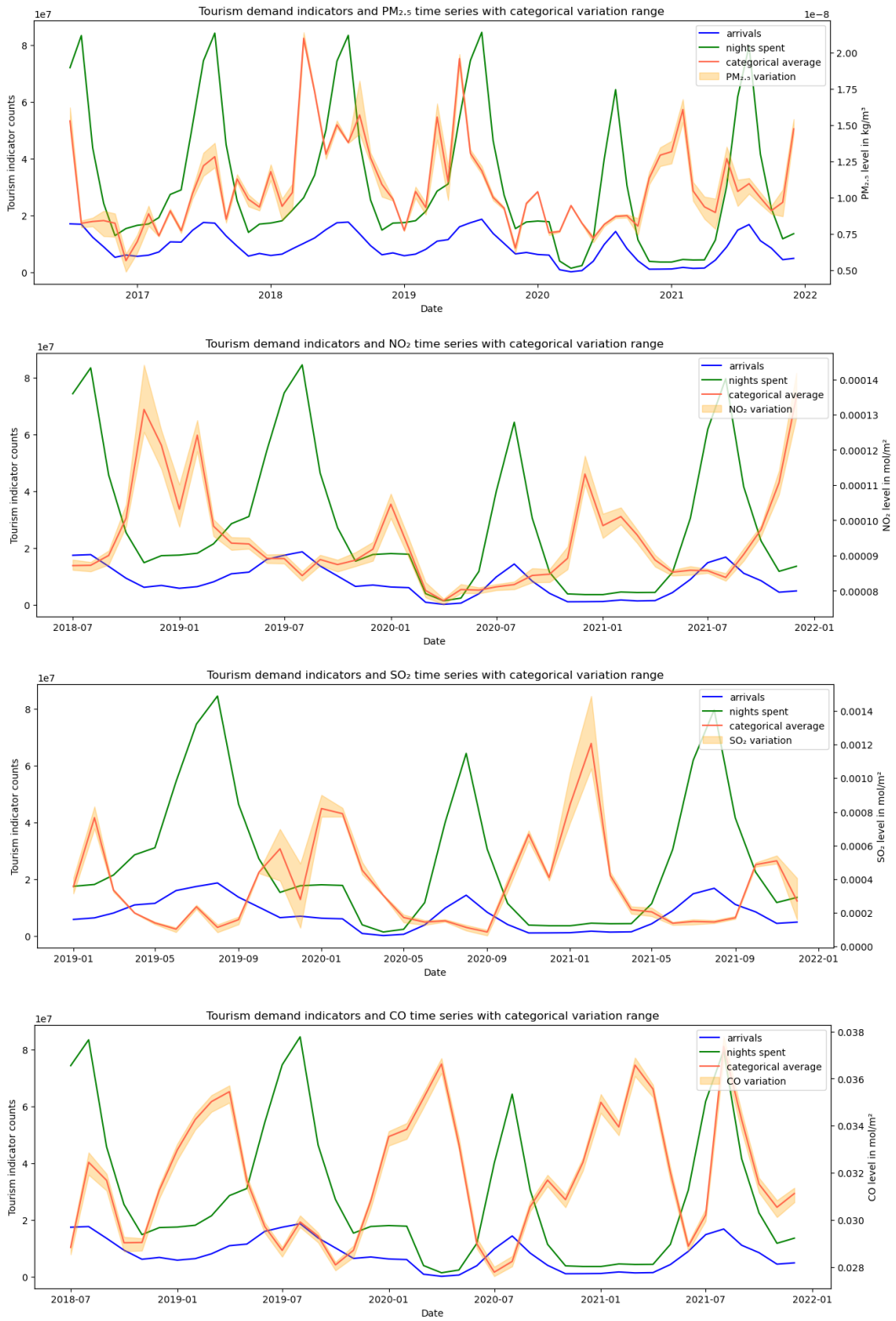

Figure 4.4: Tourism demand with Sentinel-5P and CAMS pollutant measurements showing the variation of the monthly pollutant levels across the different tourism location categories with the categorical average for each pollutant.

Pollutant	Location category of correlation	Tourist demand feature	Standard Correlation Coefficient r
$PM_{2.5}$	accommodation sites	tourist arrivals	0.2675
		tourist nights spent	0.2261
	transportation sites	tourist arrivals	0.3461
		tourist nights spent	0.3042
	tourist attractions	tourist arrivals	0.3799
		tourist nights spent	0.3381
CO	accommodation sites	tourist arrivals	-0.3953
		tourist nights spent	-0.3186
	transportation sites	tourist arrivals	-0.4015
		tourist nights spent	-0.3246
	tourist attractions	tourist arrivals	-0.3889
		tourist nights spent	-0.3159
SO_2	accommodation sites	tourist arrivals	-0.5161
		tourist nights spent	-0.5254
	transportation sites	tourist arrivals	-0.5024
		tourist nights spent	-0.5110
	tourist attractions	tourist arrivals	-0.4557
		tourist nights spent	-0.4523
NO_2	accommodation sites	tourist arrivals	-0.2267
		tourist nights spent	-0.3080
	transportation sites	tourist arrivals	-0.2110
		tourist nights spent	-0.2900
	tourist attractions	tourist arrivals	-0.2328
		tourist nights spent	-0.3077

Table 4.1: National level correlation analysis of tourism demand and pollutants at each location category.

4.3 Regional correlation analysis for tourism demand and $PM_{2.5}$ pollutant

This section focuses on regional areas to evaluate if the trends on a national level are similar on a regional level. In Table 4.2 and Table 4.3 the correlation values are given of $PM_{2.5}$ with arrivals and nights spent respectively. The regions are grouped by the divisions of the first-level Nomenclature of Territorial Units for Statistics of the European Union. For this pollutant, the same patterns are visible for the national level analysis where the correlation between the tourism arrivals are generally higher than the values for the correlation with nights spent. Across the three tourism location categories, the correlations are mostly found to be similar for both tourism demand indicators.

	Region of Correlation	**Standard Correlation Coefficient r: *tourist arrivals* and *$PM_{2.5}$* *at sites of each category***		
		Accommodation sites	Transportation sites	Tourist attraction sites
North-West	Aosta-Valley	-0.1071	-0.1520	-0.1649
	Liguria	0.4351	0.3738	0.3411
	Lombardy	0.0840	0.0184	0.0764
	Piedmont	0.0034	0.0073	0.1429
North-East	Emilia-Romagna	0.3495	0.3602	0.3405
	Friuli-Venezia Giulia	0.3801	0.3742	0.4184
	Trentino South-Tyrol	-0.0920	-0.0979	-0.0945
	Veneto	-0.0593	-0.0476	-0.0686
Center	Lazio	0.0616	0.3053	0.2779
	Tuscany	0.4908	0.4968	0.4834
	Marche	0.3673	0.3698	0.3608
	Umbria	0.4714	0.4741	0.4740
South	Abruzzo	0.3645	0.3722	0.3699
	Apulia	0.3791	0.4361	0.4452
	Basilicata	0.5775	0.5539	0.5618
	Calabria	0.4914	0.4817	0.4864
	Campania	0.4100	0.4750	0.4410
	Molise	0.3114	0.3152	0.3090
Islands	Sardinia	0.2417	0.3060	0.3445
	Sicily	0.2303	0.2326	0.2205

Table 4.2: Regional correlation analysis of tourism arrivals and $PM_{2.5}$ at each location category.

Looking at the different territorial units an observation can be made that regions in the south are showing more correlation with the $PM_{2.5}$ pollutant than other territorial units. They could be explained by the fact that most of the polluting industries are located in the north of the country. When moving to regions that are located more in the norther part it can be seen that the correlation decreases rapidly. Regions like Trentio South-Tyrol and Veneto are examples where there is almost no statistical link between the pollutant and the two tourism indicators. Furthermore, when analysing the spatial coverage of the most polluted areas in the north, one can see that the pollutant in the regions Friuli-Venezia Giulia, Emilia-Romagna and Liguria lying in the North-East and North-West parts of the country are more correlated with the tourism demand than other regions in these territorial units. Regarding these three regions in the north, in Figure 4.5 a map can be seen of Italy showing the average $PM_{2.5}$ levels in 2018 and that the regions are located on the edges of the more polluted areas above Piedmont, Lombardy and Veneto. This could indicate that the effect of the tourism demand in the regions with better overall air quality levels

	Region of Correlation	**Standard Correlation Coefficient r:** ***tourist nights spent*** **and** ***$PM_{2.5}$*** ***at sites of each category***		
		Accommodation sites	Transportation sites	Tourist attraction sites
North-West	Aosta-Valley	-0.1338	-0.1713	-0.1820
	Liguria	0.4001	0.3444	0.3142
	Lombardy	0.0869	0.0208	0.0847
	Piedmont	0.0112	0.0164	0.1585
North-East	Emilia-Romagna	0.3104	0.3242	0.3064
	Friuli-Venezia Giulia	0.3171	0.3090	0.3481
	Trentino South-Tyrol	-0.0471	-0.0584	-0.0550
	Veneto	-0.0687	-0.0620	-0.0774
Centre	Lazio	0.0824	0.3394	0.3112
	Tuscany	0.4218	0.4327	0.4206
	Marche	0.2852	0.2884	0.2779
	Umbria	0.4311	0.4326	0.4325
South	Abruzzo	0.3376	0.3410	0.3423
	Apulia	0.3054	0.3603	0.3652
	Basilicata	0.5508	0.5214	0.5320
	Calabria	0.4510	0.4406	0.4577
	Campania	0.3718	0.4404	0.4051
	Molise	0.2979	0.3014	0.2964
Islands	Sardinia	0.2191	0.2870	0.3307
	Sicily	0.1685	0.1744	0.1598

Table 4.3: Regional correlation analysis showing standard correlation coefficient of tourism nights spent and $PM_{2.5}$ at each location category.

is better measurable compared to places that have higher levels of pollution. In Italy, it is challenging for the pollutants in the northern regions to escape due to the mountainous geographical situation in the area. In the regions with more pollution, the effect of tourism could be more challenging to measure making it harder to effectively model the air quality level with the tourism demand indicators.

4.4 Granger-causality test results for $PM_{2.5}$

The national level results are first analysed showing the results in Table 4.4. Different lag times are considered to understand if the tourism demand indicators are causing an effect that can be measured later on in the air pollutant measurements. As this study is using monthly data for the modelling process, the time lags represent the number of months that are shifted in order to see if there is an effect after the fixed time lag.

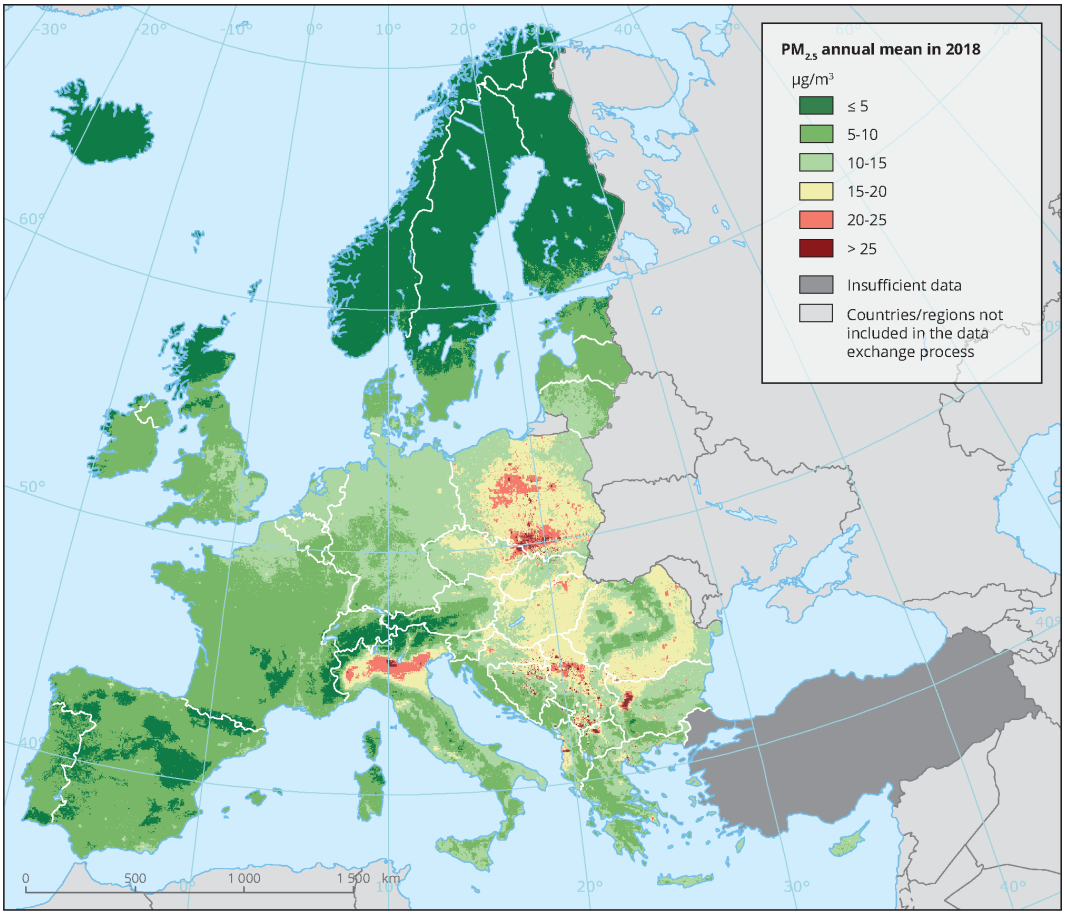

Figure 4.5: Average $PM_{2.5}$ distribution in 2018 showing high concentration of the pollutant in the northern part compared to southern regions in Italy [2].

The results for the analysis of tourism effect on air quality shows lower p-values for the first lag value suggesting a causal effect between the tourism demand time series and the air pollutant levels. Most notably, looking at nights spent, the p-values show 0.0544, 0.0458 and 0.0778 for accommodation, transportation and tourist attraction sites respectively. This suggests a high predictability of $PM_{2.5}$ levels across these locations given the time series of nights spent. For the case of the tourist arrivals, this first lag values shows less significant results. However, the correlation analysis in the previous section showed that regional values differed based on the region that was analysed. In Table 4.6 it is seen that depending on the region, arrivals is suggested to also be able to predict the $PM_{2.5}$ time series. For example, regions like Campania and Sicily show p-values of below 0.05 across all tourist categories analysed (accommodation sites, transportation sites and tourist attractions).

Interestingly, the reverse direction of the effect of air pollution on the tourism arrivals and nights spent times series shows a low likelihood from the Granger-causality test as seen

Tourism indicator	**Lag**	**Granger-causality P-value for *tourism indicator value* to predict $PM_{2.5}$ *at sites of each category***		
		Accommodation sites	Transportation sites	Tourist attraction sites
Arrivals	1	0.1369	0.1339	0.2904
	2	0.4432	0.4189	0.6797
	3	0.4745	0.4608	0.7360
	4	0.4758	0.4415	0.8142
	5	0.6347	0.5840	0.8817
Nights spent	1	0.0544	0.0458	0.0778
	2	0.3444	0.2945	0.3535
	3	0.5165	0.4890	0.6175
	4	0.4708	0.4130	0.6271
	5	0.4241	0.3681	0.4680

Table 4.4: National level Granger-causality test results to see if tourist arrivals and nights spent can predict $PM_{2.5}$ levels showing the p-value for 5 lagged months for each category.

Tourism indicator	**Lag**	**Granger-causality P-value for $PM_{2.5}$ *at sites of each category* to predict *tourism indicator value***		
		Accommodation sites	Transportation sites	Tourist attraction sites
Arrivals	1	0.5915	0.8854	0.7640
	2	0.6168	0.5072	0.4632
	3	0.8538	0.7560	0.6362
	4	0.8769	0.7985	0.6997
	5	0.3921	0.3286	0.4171
Nights spent	1	0.3964	0.5837	0.9968
	2	0.5245	0.4800	0.4503
	3	0.7820	0.7362	0.5828
	4	0.9056	0.8890	0.7743
	5	0.8663	0.8585	0.7653

Table 4.5: National level Granger-causality test results to see if $PM_{2.5}$ levels can predict tourist arrivals and nights spent showing the p-value for 5 lagged months for each category.

in Table 4.5. This can be seen for instance in the first row of the Granger-test results between the $PM_{2.5}$ pollutant and nights spent, where the p-values are showing 0.3964, 0.5837 and 0.9968 for accommodation, transportation and tourist attraction sites respectively.

Besides giving information on how well the time series are fit to model each other in terms of Granger-causality, the results for the different lag times give insight in effects of tourism on air quality. Observing when a change in the tourist arrivals and nights spent time series affects the $PM_{2.5}$ levels in different regions could show potential to model the

air quality based on these tourism indicators. Overall it is seen in the tables of the Granger-causality analysis that the first lag value is often a more significant predictor for the $PM_{2.5}$ levels. It is expected that regions which show more significant p-values would lend themselves to modelling better in time series models. The information of the first lag value being more significant is used in the next section to model $PM_{2.5}$ levels using the tourism demand indicators with an LSTM model.

For the regional level analysis Table 4.6 and Table 4.7 show the Granger statistics produced for analysing the tourism arrivals and nights spent to predict the $PM_{2.5}$ pollutant. Similar findings are found for the p-values in regions with overall lower levels of $PM_{2.5}$ like in the case of correlation analysis. The finding that the p-value tends to be more significant for the first lag value across the three categories holds for both arrivals and nights spent. This indicates the potential to use tourism arrivals and nights spent data from one month to predict the $PM_{2.5}$ levels of the next month.

Similarly, making the comparison across the different tourism category locations (i.e. comparing the columns of the table), it can be noted that predictability of $PM_{2.5}$ at tourist attraction sites can deviate from the other two categories. For example, in Table 4.6 which uses the arrivals time series, Calabria shows that the predictability of $PM_{2.5}$ at the tourist attraction sites is significantly lower showing a p-value of 0.1706 compared to 0.0447 for accommodation sites and 0.0640 for transportation sites where lag value is 1. This findings is also seen when using the nights spent time series in Table 4.7 for the case of Calabria.

4.5 LSTM time series modelling of $PM_{2.5}$ levels using tourism demand indicators

In this section the LSTM modelling process of the $PM_{2.5}$ pollutant using the tourism demand indicators as input features is presented. The pollutant levels at the three categories of tourism locations will be modelled using the tourism demand indicators which are arrivals, nights spent and average length of stay. First, a national level analysis is performed and the results of this are presented in terms of model configuration, modelling performance and hyperparameter tuning. Afterwards, the same study is conducted on a regional level to see how well the results of the two spatial levels relate to each other and to see how the correlation and Granger-causality findings for each region are reflected in terms of modelling behaviour.

4.5.1 LSTM input features and hyperparameter tuning

The LSTM analysis that is performed in this research is designed such that the tourism arrivals, nights spent, and average length of stay are used as input features to model the $PM_{2.5}$ levels at the three categories of tourism sites. The monthly data about the national and regional tourism demand features and the $PM_{2.5}$ levels across the three location categories span a time period over 5 years and 6 months. The training set is created by using 70% of the available $PM_{2.5}$ data and the validation and test set constitute to the remaining data, each using 15%. To implement this model using the Keras library, a number of parameters

	Region of Granger Causality Test	**Lag**	**Granger-causality P-value for *tourist arrivals* to predict $PM_{2.5}$ *at sites of each category***		
			Accommodation sites	Transportation sites	Tourist attractions
North-West	Aosta-Valley	1	0.5301	0.4176	0.3879
		2	0.2659	0.1908	0.1724
	Liguria	1	0.3622	0.4791	0.5507
		2	0.3960	0.4758	0.5313
	Lombardy	1	0.5532	0.4660	0.5209
		2	0.8800	0.7783	0.8347
	Piedmont	1	0.3055	0.3674	0.7267
		2	0.5032	0.5565	0.7406
North-East	Emilia-Romagna	1	0.3803	0.4139	0.4544
		2	0.5716	0.6529	0.7049
	Friuli-Venezia Giulia	1	0.0232	0.0306	0.0176
		2	0.2348	0.2855	0.1702
	Trentino South-Tyrol	1	0.9506	0.9100	0.9387
		2	0.6894	0.6979	0.7236
	Veneto	1	0.9271	0.9249	0.9509
		2	0.5786	0.6500	0.5712
Centre	Lazio	1	0.1968	0.1073	0.1127
		2	0.4606	0.2740	0.2895
	Tuscany	1	0.3566	0.3863	0.4978
		2	0.2417	0.2960	0.3607
	Marche	1	0.0252	0.0235	0.0192
		2	0.0174	0.0210	0.0240
	Umbria	1	0.1476	0.1529	0.1546
		2	0.2419	0.2250	0.2249
South	Abruzzo	1	0.0360	0.0582	0.0474
		2	0.0176	0.0148	0.0237
	Apulia	1	0.0035	0.0045	0.0066
		2	0.0078	0.0038	0.0028
	Basilicata	1	0.0057	0.0011	0.0015
		2	0.0185	0.0048	0.0068
	Calabria	1	0.0447	0.0640	0.1706
		2	0.0008	0.0013	0.0048
	Campania	1	0.0140	0.0161	0.0173
		2	0.1429	0.1401	0.1481
	Molise	1	0.0906	0.0996	0.1121
		2	0.2535	0.2708	0.3161
Islands	Sardinia	1	0.0685	0.0625	0.0659
		2	0.3673	0.4697	0.4831
	Sicily	1	0.0134	0.0181	0.0232
		2	0.0329	0.0464	0.0484

Table 4.6: Regional Granger-causality analysis showing p-values for a lag value of one and two months for tourism arrivals and $PM_{2.5}$ at each location category.

need to be set. This study looks into learning rate, number of units, batch size and number of epochs for the learning. The convergence speed of the model is also determined by the learning rate and in the case of a too high learning rate it can potentially causing unstable

	Region of Granger Causality Test	**Lag**	**Granger-causality P-value for *tourist nights spent* to predict *$PM_{2.5}$ at sites of each category***		
			Accommodation sites	Transportation sites	Tourist attractions
North-West	Aosta-Valley	1	0.5553	0.4443	0.4146
		2	0.1252	0.0902	0.0820
	Liguria	1	0.0606	0.1014	0.1315
		2	0.4752	0.6801	0.7893
	Lombardy	1	0.9705	0.8512	0.9886
		2	0.6541	0.5762	0.5919
	Piedmont	1	0.7700	0.8766	0.6655
		2	0.3603	0.3192	0.3235
North-East	Emilia-Romagna	1	0.0915	0.0971	0.1144
		2	0.4027	0.4521	0.5184
	Friuli-Venezia Giulia	1	0.0097	0.0134	0.0076
		2	0.0501	0.0812	0.0395
	Trentino South-Tyrol	1	0.3783	0.4193	0.4303
		2	0.3639	0.3635	0.3704
	Veneto	1	0.6290	0.6258	0.6586
		2	0.8949	0.8784	0.8794
Centre	Lazio	1	0.1266	0.0890	0.0867
		2	0.6639	0.5354	0.5309
	Tuscany	1	0.0581	0.0655	0.1029
		2	0.2211	0.2239	0.2492
	Marche	1	0.0091	0.0081	0.0066
		2	0.0193	0.0221	0.0241
	Umbria	1	0.0452	0.0515	0.0526
		2	0.1377	0.1280	0.1283
South	Abruzzo	1	0.0126	0.0223	0.0168
		2	0.0095	0.0097	0.0138
	Apulia	1	0.0104	0.0101	0.0126
		2	0.0217	0.0125	0.0095
	Basilicata	1	0.0062	0.0009	0.0012
		2	0.0181	0.0035	0.0054
	Calabria	1	0.0533	0.0672	0.1576
		2	0.0042	0.0068	0.0167
	Campania	1	0.0017	0.0014	0.0018
		2	0.0489	0.0398	0.0500
	Molise	1	0.0258	0.0287	0.0310
		2	0.1012	0.1085	0.1160
Islands	Sardinia	1	0.0563	0.0524	0.0565
		2	0.3576	0.4176	0.4219
	Sicily	1	0.0176	0.0255	0.0286
		2	0.0314	0.0503	0.0545

Table 4.7: Regional Granger-causality analysis showing p-values for a lag value of one and two months for tourism nights spent and $PM_{2.5}$ at each location category.

results. Through manual tuning, the learning rate was set to 0.01 considering the outputs being more in line with the shape of the modelled pollutants. Considering the input data to be monthly data, using the indicated learning rate the aim is to achieve higher generalization

performance without underfitting to lose the monthly tourism trends in the input features.

By using the tourism demand indicators as the input features, the goal is that the LSTM model captures the seasonal nature of the tourism flow to model the values of the $PM_{2.5}$ pollutant levels at the three categorical sites. For preparation of the dataset for the learning on the tourism indicators and corresponding pollutant levels, the data needs to be prepared using the sliding window approach such that clear input-output pairs are created from the input sequence which is required for the used LSTM model. In this supervised learning approach, local patterns in the input sequence can then be used such that the model can capture temporal patterns between the number of tourist arrivals, the nights they spent and the average length of stay with the pollutant levels. From the Granger-causality test results, it was found that the arrivals and nights spent indicators where mostly significant for the first lag value suggesting a high predictability of the $PM_{2.5}$ pollutant level using the tourism arrivals and nights spent indicators form the month before. This finding is used to prepare the input sequence using this lag value for the sliding window indicating that the model learns the pollutant levels at three tourism categories using the tourist demand data from the previous month. While a longer window size could provide more contextual information regarding long-term dependencies between the tourism demand features used, the information about the high predictability of the $PM_{2.5}$ pollutant using the tourism demand values of the previous month could be lost. This granularity for the input sequence therefore aims to capture the short-term fluctuations of the tourism demand features and its response on the $PM_{2.5}$ levels to aim for higher predictability of the pollutant the three categories of tourism locations.

Grid search is used to find optimal values for the number of units and batch size using early stopping for the number of epochs when the validation error does not decrease for 5 epochs to prevent overfitting on the training set. The values that are considered for the batch size, indicating the number of prepared training input samples processed in one pass through the LSTM model, are 4, 8, 16. It was chosen to use these relatively smaller batch sizes considering the size of the available training data for the modelling. Using a smaller batch size can help to increase the utilization of data by having the model update its weights more frequently. Although a smaller batch size would increase the computational demand of the model, making use of a smaller batch size can help lowering the risk of overfitting considering the monthly granularity of the input data by having more variability in updating the weights after each pass through the model.

The number of unites that are considered by the grid search process are 8, 16, 32. In the context of the LSTM model, the number of units determine the capacity of each of the layer in the sequentially stacked layers. The input units process the input training data to retain the relevant information it learns about the relation between the tourism indicators and corresponding $PM_{2.5}$ levels. With a higher number of unites the model can learn detailed patterns in the data. In the next section, the results of the hyperparameter tuning together with the RMSE scores for the national and regional level modelling across the different location categories are discussed.

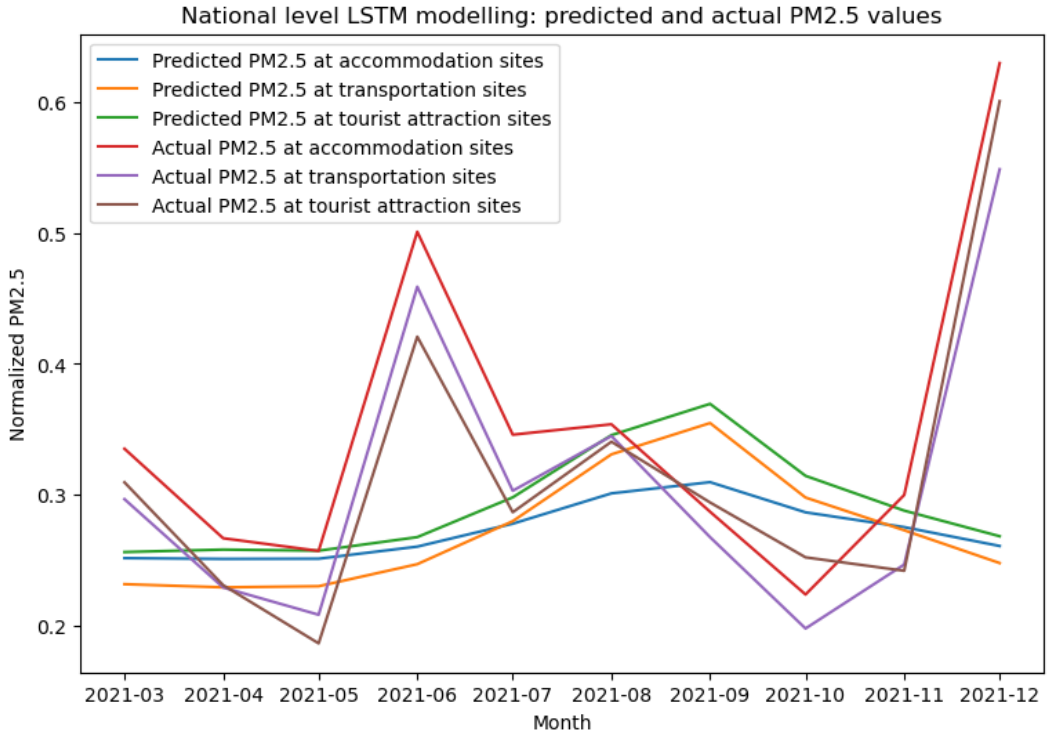

Figure 4.6: LSTM time series modelling of $PM_{2.5}$ pollutant levels using tourism demand indicators using national level data.

4.5.2 LSTM modelling results of $PM_{2.5}$ based on tourism demand indicators on a national and regional level

Performing the grid search optimization for the input tourism data and $PM_{2.5}$ pollutant levels on a national level resulted in a RMSE of 0.1461, 0.1261 and 0.1242 for modelling $PM_{2.5}$ at the accommodation, transportation and tourist attraction sites respectively. The model makes use of a batch size of 4, layers with 16 units and using 9 epochs during the training process. In Figure 4.6 the predictions of the model for each location category can be found suggesting the aim of the model to a capturing the increase in pollutants during the summer season. The same analysis is performed on a regional level with the results of the hyperparameter tuning given in Table 4.8. Looking at the RMSE values of the regional modelling, the error scores of the national level modelling are about average when looking at range of error scores of the regional level modelling. The batch size and the number of units are varying across the regions and different parts of the country. A pattern that is noticeable is the relatively higher number of epochs for the regions in the north, suggesting a higher effort needed to effectively model these regions. This observation is in line with the correlation analysis and the Granger-causality testing from the previous sections indicating that in higher polluted regions like in the north, it is harder to derive clear trends and causal relations between tourism arrivals and nights spent and the pollutants. Regions that

	Region	Accommodation sites RMSE	Transportation sites RMSE	Tourist attraction sites RMSE	Num units	Batch size	Num epochs
North-West	Aosta-Valley	0.1311	0.1304	0.1328	8	4	24
	Liguria	0.1609	0.1795	0.1873	8	8	10
	Lombardy	0.2717	0.3045	0.2729	8	8	17
	Piedmont	0.1916	0.2439	0.2175	16	16	16
North-East	Emilia-Romagna	0.1574	0.1520	0.1644	16	16	18
	Friuli-Venezia Giulia	0.1648	0.1279	0.1131	16	4	14
	Trentino South-Tyrol	0.1706	0.1748	0.1762	32	16	81
	Veneto	0.3414	0.3244	0.3526	8	16	32
Centre	Lazio	0.2503	0.1584	0.1754	16	4	7
	Tuscany	0.1005	0.1074	0.0931	32	4	19
	Marche	0.1031	0.1048	0.1215	32	4	16
	Umbria	0.1098	0.0743	0.0737	8	4	6
South	Abruzzo	0.1993	0.1200	0.1324	16	4	5
	Apulia	0.2317	0.1882	0.1482	16	8	6
	Basilicata	0.1159	0.1245	0.1044	16	8	5
	Calabria	0.2931	0.2158	0.1325	32	16	6
	Campania	0.2250	0.1938	0.2025	16	8	11
	Molise	0.1007	0.1076	0.0850	16	4	5
Islands	Sardinia	0.2062	0.1733	0.1768	16	8	8
	Sicily	0.2001	0.1928	0.1623	8	16	8

Table 4.8: LSTM modelling of regional $PM_{2.5}$ showing RMSE for each tourism location category together with the results of the hyperparameter tuning.

are found to be less correlated and indicate lower predictability through Granger-causality testing like Lombardy and Veneto also show higher error scores across the three categories. In Table 4.8, note the RMSE scores of 0.2717, 0.3045 and 0.2729 for accommodation, transportation and attraction sites respectively for Lombardy. Lombardy is a north-western region with a strong industrial footprint and mountainous natural setting, thus trapping pollutants in its atmosphere. Similarly, for Veneto these are 0.3414, 0.3244 and 0.3526 with comparable characteristics to Lombardy. This observation adds to the reasoning that the effect of tourism on air quality can better be modelled in regions with cleaner overall air quality.

Chapter 5

Conclusions and Future Work

This chapter gives an overview of the research that was performed together with its contributions. Following this overview is a reflection on the work where the process of coming to the contributions is discussed. Finally, this work ends with ideas for future research.

5.1 Contributions

In this work, data integration was performed through data retrieval and pre-processing to create the datasets needed to enable the modelling of the effects of tourism demand on air quality. This gives researchers and practitioners aiming to perform similar data integration efforts a general guide on how this could be performed. Presenting the challenges of the data integration process in the methodology where the data pipeline served as a workflow contributes to the understanding around the challenges of pre-processing and joining the data that is needed for effective modelling. In this study, the challenges that were encountered when using data from administrative divisions during the data retrieval phase could also arise in a general sense for similar data retrieval and integration tasks. Depending on how administrative divisions are structured and the potentially deviating methods they employ on collecting and present their data, more steps in retrieval process could be introduced resulting in additional pre-processing practices. Knowledge about the case at hand could provide insights on how the pre-processing can be performed best in such scenarios.

The study analyses three different categories of spatial sites important for the tourism industry to broadly model the effects of tourism on the air quality in these locations. The different stages of the data modelling process gave insights into how tourism locations can be retrieved through using an RDF based knowledge base. This provided categorised location data for sampling the different kinds of air pollutants in this study. This data was then used for correlation analysis between categorised locations and the tourism demand indicators which were arrivals and nights spent. From this analysis, $PM_{2.5}$ was found the be more correlated with tourism demand features compared to the pollutants from the Sentinel-5P product. Furthermore, the regional analysis of the $PM_{2.5}$ pollutant found that regions with overall less $PM_{2.5}$ show higher correlations with the tourism demand features compared to regions with higher levels of $PM_{2.5}$.

The analysis was continued using Granger-causality testing to see if the tourism demand indicators can predict the values for the $PM_{2.5}$ pollutant levels at the different categories of locations in the different regions. The results suggest that using a lag value of one month could be used to predict the $PM_{2.5}$ pollutant, indicating that the time series for tourism arrivals and nights spent could potentially be used to predict the $PM_{2.5}$ time series for the following month.

Thus, there were significant differences between the regions, with regions in the north showing much lower values for the correlation compared to regions in the south. When looking at the overall $PM_{2.5}$ distribution in Italy, it can be seen that these regions are mostly overlapping with the strongly polluted northern parts of the country. Performing Granger-causality testing also showed this pattern with southern regions having notably lower p-values. An LSTM model was eventually trained using the information of this lagged relationship. National level LSTM modelling resulted in a RMSE which was similar to the average of the RMSE for the regional analysis. It was found that regions with a high overall $PM_{2.5}$ are challenging to model and spend more epochs for modelling the effects of tourism demand on air quality. For the case of $PM_{2.5}$, the modelling in this work suggests that the effect of tourism demand can better be modelled in regions with cleaner air quality with respect to this pollutant.

5.2 Conclusions

This research performed multi-source data modelling to understand the effects of tourism demand on air quality. Correlation analysis and Granger-causality testing were used in this study to better understand the relation between the tourism demand and pollutant levels across accommodation, transportation and tourism attractions. These locations were chosen to broadly model the effect of tourism on the air quality given their importance for the tourism industry. In the end, an LSTM model was trained to understand how well one can predict $PM_{2.5}$ levels at the three categories of tourism locations.

To enable this modelling, multi-source data was combined which was done following a data processing pipeline. When retrieving tourism demand data on different administrative levels, challenges can arise that require additional pre-processing to effectively use the data for the analysis like in the case of Trentino-Alto Adige/Südtirol in this study. Correlation analysis indicates a higher correlation between the tourism demand indicators and $PM_{2.5}$ in overall cleaner regions for this pollutant. In these regions, Granger-causality testing using different lags to analyse causality suggests a higher chance of predictability of the $PM_{2.5}$ time series using tourism demand data from the previous month. Training an LSTM model using the information of this lagged relationship for each region suggests that regions with a high overall $PM_{2.5}$ are challenging to model showing more epochs being performed to learn the effects of tourism demand on air quality. This is in line with the findings for the $PM_{2.5}$ pollutant being lower correlated with tourism features and showing higher p-values for the Granger-causality testing for the regions with overall high $PM_{2.5}$ levels. It is therefore suggested that analysis in this field of sustainable tourism could gain a better understanding of the effects of tourism demand when measuring less overall polluted regions.

5.3 Discussion/Reflection

For this analysis, the format of the tourism statistics on arrivals and nights spent determined the choice of the temporal granularity for the air quality sampling in GEE. The air quality samples where aggregated to monthly measurement to match the temporal granularity of the statistics that were available through ISTAT. Potential future availability of data on a finer scale could provide more detailed analysis which could also lead to learning more detailed relationships between the tourism demand indicators and the sampled pollutants by the LSTM model.

The higher correlation of the tourism demand features with $PM_{2.5}$ compared to the findings for NO_2, SO_2 and CO could be explained by the fact the CAMS sampling uses in situ data for the $PM_{2.5}$ pollutant which provides more detailed measurements. Based on this, it could be the case that measuring the effects of tourism on air quality using Sentinel-5P data is more challenging compared to using data from CAMS. From the analysis and plotting of the pollutants sampled with the Sentinel-5P product, the peaks of these pollutants where mostly visible when the tourist arrivals and nights spent were low. Only since July of 2021, GEE started to provide in situ data using CAMS data for NO_2, SO_2 and CO. The pollutants could potentially be modelled more reliably through the model introduced in this research paper when there becomes more in situ data available for these pollutants from the GEE platform.

As the LSTM model performs better in regions which do not show high $PM_{2.5}$ levels, the applicability of this research is higher for tourist regions with cleaner overall air quality, as measured by $PM_{2.5}$. Considering the data availability for this pollutant, it is expected that the modelling of air quality would benefit from more data leading to better predictability also in cleaner regions. Without the benefit of extensive data analysis for these countries, looking at Figure 4.5 one could think of Portugal, Scotland, Ireland and other similar tourist destinations, benefiting from the research presented here to continue to monitor the air quality of their tourist regions while continuing to benefit economically from the tourist industry.

5.4 Future work

The data integration process in the work could be applied to different countries to see how well the results found in this work can be generalized to other study areas. It would be interesting to observe if other countries show similar Granger-causality results for the different lag values that were analysed. The analysis could be expanded to different types of machine learning models as well to compare the performance of different types of models.

Looking at places with different levels of overall air pollution could also provide insights to understand if similar challenges are found in other study areas in terms of modelling air pollutants using tourism demand indicators across different regions.

Over time when transport facilities become more environmentally friendly, future work can analyse if the patters across accommodation, transportation and tourist attractions are still the same as found in this study. Similar studies can also look at the effects of tourism campaigns with the aim of achieving a more equal distribution of tourists across a study area and seeing how the air quality levels respond as a result.

Bibliography

[1] Tomiwa Sunday Adebayo, Seyi Saint Akadiri, Obioma Chinenyenwa Asuzu, Nanfa Hamisu Pennap, and Yetunde Sadiq-Bamgbopa. Impact of tourist arrivals on environmental quality: a way towards environmental sustainability targets. *Current Issues in Tourism*, 26(6):958–976, 2023.

[2] European Environment Agency. PM2.5 annual mean in 2018 , 2020. URL https://www.eea.europa.eu/data-and-maps/figures/pm2-5-annual-mean-in-2.

[3] Fayyaz Ahmad, Muhammad Draz, Lijuan Su, Ilhan Ozturk, and Abdul Rauf. Tourism and Environmental Pollution: Evidence from the One Belt One Road Provinces of Western China. *Sustainability*, 10(10):3520, Sep 2018. ISSN 2071-1050. doi: 10.3390/su10103520. URL http://dx.doi.org/10.3390/su10103520.

[4] Najid Ahmad and Xuejiao Ma. How does tourism development affect environmental pollution? *Tourism Economics*, 28(6):1453–1479, 2022.

[5] A. Azapagic and S. Perdan. Indicators of Sustainable Development for Industry: A General Framework. *Process Safety and Environmental Protection*, 78(4):243–261, 2000. ISSN 0957-5820. doi: https://doi.org/10.1205/095758200530763. URL https://www.sciencedirect.com/science/article/pii/S0957582000708834. Sustainable Development.

[6] The World Bank. International tourism, number of arrivals - Italy. URL https://data.worldbank.org/indicator/ST.INT.ARVL?locations=IT.

[7] Bianca Biagi and Manuela Pulina. Bivariate VAR models to test Granger causality between tourist demand and supply: Implications for regional sustainable growth. *Papers in Regional Science*, 88(1):231–244, 2009.

[8] Richard W. Butler. Tourism, Environment, and Sustainable Development. *Environmental Conservation*, 18(3):201–209, 1991. doi: 10.1017/S0376892900022104.

[9] Rosaria Rita Canale and Rita De Siano. Territorial pressure and tourism contribution to GDP: The case of Italian regions. *International Journal of Tourism Research*, 23 (5):891–900, 2021.

[10] Mercedes Castro-Nuño, José A Molina-Toucedo, and Maria P Pablo-Romero. Tourism and GDP: A meta-analysis of panel data studies. *Journal of Travel research*, 52(6): 745–758, 2013.

[11] J Chaplin and Lars Brabyn. Using remote sensing and GIS to investigate the impacts of tourism on forest cover in the Annapurna Conservation Area, Nepal. *Applied Geography*, 43:159–168, 2013.

[12] Prem Chhetri and Colin Arrowsmith. GIS-based Modelling of Recreational Potential of Nature-Based Tourist Destinations. *Tourism Geographies - TOUR GEOGR*, 10: 233–257, 04 2008. doi: 10.1080/14616680802000089.

[13] Lixia Chu, Francis Oloo, Helena Bergstedt, and Thomas Blaschke. Assessing the Link between Human Modification and Changes in Land Surface Temperature in Hainan, China Using Image Archives from Google Earth Engine. *Remote Sensing*, 12(5):888, Mar 2020. ISSN 2072-4292. doi: 10.3390/rs12050888. URL http://dx.doi.org/10.3390/rs12050888.

[14] Lixia Chu, Francis Oloo, Bin Chen, Miaomiao Xie, and Thomas Blaschke. Assessing the Influence of Tourism-Driven Activities on Environmental Variables on Hainan Island, China. *Remote Sensing*, 12:2813, 08 2020. doi: 10.3390/rs12172813.

[15] Sefa Awaworyi Churchill, Lei Pan, and Sudharshan Reddy Paramati. Air pollution and tourism: Evidence from G20 countries. *Journal of Travel Research*, 61(2):223–234, 2022.

[16] Oscar Saenz de Miera and Jaume Rosselló. Modeling tourism impacts on air pollution: The case study of PM10 in Mallorca. *Tourism Management*, 40:273–281, 2014. ISSN 0261-5177. doi: https://doi.org/10.1016/j.tourman.2013.06.012. URL https://www.sciencedirect.com/science/article/pii/S0261517713001313.

[17] Glauco De Vita and Khine S Kyaw. Tourism development and growth. *Annals of Tourism Research*, 60, 2016.

[18] Fumin Deng, Yuan Fang, Lin Xu, and Zhi Li. Tourism, transportation and low-carbon city system coupling coordination degree: A case study in Chongqing Municipality, China. *International Journal of Environmental Research and Public Health*, 17(3): 792, 2020.

[19] Celeste Eusébio, Vitor Rodrigues, Maria João Carneiro, Mara Madaleno, Margarita Robaina, and Alexandra Monteiro. The role of air quality for reaching tourism environmental sustainability: A segmentation approach based on visitors' pro-environmental behaviors. *International Journal of Tourism Research*, 2023.

[20] Douglas C Frechtling. Assessing the Impacts of Travel and Tourism-Measuring Economic Benefits'. *International Library of Critical Writings in Economics*, 121:9–27, 2000.

[21] Andy Hamilton, Hongxia Wang, Ali Murat Tanyer, Yusuf Arayici, Xiaonan Zhang, and Yonghui Song. Urban information model for city planning. *Journal of Information Technology in Construction*, 10:55–67, 2005.

[22] Anne Hardy, Robert JS Beeton, and Leonie Pearson. Sustainable tourism: An overview of the concept and its position in relation to conceptualisations of tourism. *Journal of sustainable tourism*, 10(6):475–496, 2002.

[23] Salah S Hassan. Determinants of market competitiveness in an environmentally sustainable tourism industry. *Journal of travel research*, 38(3):239–245, 2000.

[24] James ES Higham. *Critical issues in ecotourism: Understanding a complex tourism phenomenon*. Routledge, 2007.

[25] Muhammad Irfan, Sami Ullah, Asif Razzaq, Jinyang Cai, and Tomiwa Sunday Adebayo. Unleashing the dynamic impact of tourism industry on energy consumption, economic output, and environmental quality in China: A way forward towards environmental sustainability. *Journal of Cleaner Production*, 387:135778, 2023. ISSN 0959-6526. doi: https://doi.org/10.1016/j.jclepro.2022.135778. URL https://www.sciencedirect.com/science/article/pii/S0959652622053525.

[26] Stanislav Ivanov and Craig Webster. Measuring the impact of tourism on economic growth. *Tourism economics*, 13(3):379–388, 2007.

[27] Salih Turan Katircioglu. International tourism, energy consumption, and environmental pollution: The case of Turkey. *Renewable and Sustainable Energy Reviews*, 36:180–187, 2014. ISSN 1364-0321. doi: https://doi.org/10.1016/j.rser.2014.04.058. URL https://www.sciencedirect.com/science/article/pii/S1364032114002962.

[28] Haejin Lee, Jennifer Baylon Verances, and Woongang Song. The tourism-environment causality. *International Journal of Tourism Sciences*, 9(3):39–48, 2009.

[29] YiFei Li and Han Cao. Prediction for tourism flow based on LSTM neural network. *Procedia Computer Science*, 129:277–283, 2018.

[30] Jianjun Liu, Haili Pan, and Zheng Shiyong. Tourism Development, Environment and Policies: Differences between Domestic and International Tourists. *Sustainability*, 11:1390, 03 2019. doi: 10.3390/su11051390.

[31] Zhen Liu, Jing Lan, Fengsheng Chien, Muhammad Sadiq, and Muhammad Atif Nawaz. Role of tourism development in environmental degradation: A step towards emission reduction. *Journal of environmental management*, 303:114078, 2022.

[32] Daniel Lorente and Nuno Carlos Leitão. The role of tourism, trade, renewable energy use and carbon dioxide emissions on economic growth: evidence of tourism-led growth hypothesis in EU-28. *Environmental Science and Pollution Research*, 27, 12 2020. doi: 10.1007/s11356-020-10375-1.

[33] Tuan Hock Ng, Chun-Teck Lye, and Ying San Lim. A decomposition analysis of CO2 emissions: Evidence from Malaysias tourism industry. *International Journal of Sustainable Development & World Ecology*, 23:1–12, 12 2015. doi: 10.1080/13504509.2015.1117534.

[34] Italian National Institute of Statistics. Monthly Demographic Balance, 2022. URL https://demo.istat.it/app/?i=D7B&a=2022&l=en. Accessed: 06-12-2022.

[35] Angelo Presenza and Maria Cipollina. Analysing tourism stakeholders networks. *Tourism Review*, 2010.

[36] Copernicus Programme. About Copernicus. URL https://insitu.copernicus.eu/about.

[37] Chen Qian, Weifeng Li, Zhengyu Duan, Dongyuan Yang, and Bin Ran. Using mobile phone data to determine spatial correlations between tourism facilities. *Journal of Transport Geography*, 92:103018, 2021. ISSN 0966-6923. doi: https://doi.org/10.1016/j.jtrangeo.2021.103018. URL https://www.sciencedirect.com/science/article/pii/S0966692321000715.

[38] Paulo Ribeiro and José FG Mendes. Route planning for soft modes of transport: healthy routes. *WIT Transactions on the Built Environment*, 116:677–688, 2011.

[39] M Robaina, M Madaleno, S Silva, C Eusébio, MJ Carneiro, C Gama, K Oliveira, MA Russo, and A Monteiro. The relationship between tourism and air quality in five European countries. *Economic Analysis and Policy*, 67:261–272, 2020.

[40] Ignacio Ruiz-Guerra, Valentín Molina-Moreno, Francisco J. Cortés-García, and Pedro Núñez-Cacho. Prediction of the impact on air quality of the cities receiving cruise tourism: the case of the Port of Barcelona. *Heliyon*, 5(3):e01280, 2019. ISSN 2405-8440. doi: https://doi.org/10.1016/j.heliyon.2019.e01280. URL https://www.sciencedirect.com/science/article/pii/S2405844018373171.

[41] Oscar Saenz de Miera Berglind and Jaume Rosselló. Tropospheric ozone, air pollution and tourism: a case study of Mallorca. *Journal of Sustainable Tourism*, 21:1232–1243, 11 2013. doi: 10.1080/09669582.2013.776061.

[42] Yerik Afrianto Singgalen. Tourism infrastructure development and transformation of vegetation index in Dodola Island of Morotai Island Regency. *Journal of Information Systems and Informatics*, 4(1):130–144, 2022.

[43] Statista. Share of travel and tourism's total contribution to GDP in Italy from 2019 to 2021, 6 2022. URL https://www.statista.com/statistics/628849/tourism-total-contribution-to-gdp-italy-share/. Accessed: 28-02-2023.

[44] Umut Türk, John Östh, Karima Kourtit, and Peter Nijkamp. The path of least resistance explaining tourist mobility patterns in destination areas using Airbnb data. *Journal of Transport Geography*, 94:103130, 2021. ISSN 0966-6923. doi: https://doi.org/10.1016/j.jtrangeo.2021.103130. URL https://www.sciencedirect.com/science/article/pii/S0966692321001836.

[45] Denny Vrandečić and Markus Krötzsch. Wikidata: A Free Collaborative Knowledgebase. *Commun. ACM*, 57(10):78–85, sep 2014. ISSN 0001-0782. doi: 10.1145/2629489. URL https://doi.org/10.1145/2629489.

[46] Jieyong Wang and Yansui Liu. Tourism-Led Land-Use Changes and their Environmental Effects in the Southern Coastal Region of Hainan Island, China. *Journal of Coastal Research*, 29:1118–1125, 09 2013. doi: 10.2112/JCOASTRES-D-12-00039.

[47] Pu Wu, Yuanjun Han, and Mi Tian. The measurement and comparative study of carbon dioxide emissions from tourism in typical provinces in China. *Acta Ecologica Sinica*, 35(6):184–190, 2015. ISSN 1872-2032. doi: https://doi.org/10.1016/j.chnaes.2015.09.004. URL https://www.sciencedirect.com/science/article/pii/S1872203215000487.

[48] Chao Xiong, Asif Khan, Sughra Bibi, Hizar Hayat, and Shaoping Jiang. Tourism subindustry level environmental impacts in the US. *Current Issues in Tourism*, pages 1–19, 2022.

[49] Yang Xu, Jiaying Xue, Sangwon Park, and Yang Yue. Towards a multidimensional view of tourist mobility patterns in cities: A mobile phone data perspective. *Computers, Environment and Urban Systems*, 86:101593, 2021. ISSN 0198-9715. doi: https://doi.org/10.1016/j.compenvurbsys.2020.101593. URL https://www.sciencedirect.com/science/article/pii/S0198971520303264.

[50] Hong Yuan, Kun-xi Nie, and Xiao-ya Xu. Relationship between tourism number and air quality by carbon footprint measurement: a case study of Jiuzhaigou Scenic Area. *Environmental Science and Pollution Research*, 28, 04 2021. doi: 10.1007/s11356-020-12068-1.

[51] Fen Zhang, Haochen Peng, Xiaofan Sun, and Tianyi Song. Influence of Tourism Economy on Air Quality—An Empirical Analysis Based on Panel Data of 102 Cities in China. *International Journal of Environmental Research and Public Health*, 19 (7):4393, Apr 2022. ISSN 1660-4601. doi: 10.3390/ijerph19074393. URL http://dx.doi.org/10.3390/ijerph19074393.

[52] Lei Zhang and Jing Gao. Exploring the effects of international tourism on China's economic growth, energy consumption and environmental pollution: Evidence from

a regional panel analysis. *Renewable and Sustainable Energy Reviews*, 53:225–234, 2016. ISSN 1364-0321. doi: https://doi.org/10.1016/j.rser.2015.08.040. URL https://www.sciencedirect.com/science/article/pii/S1364032115009004.

[53] Bahram Zikirya, Jieyu Wang, and Chunshan Zhou. The relationship between CO2 emissions, air pollution, and tourism flows in China: A panel data analysis of Chinese Provinces. *Sustainability*, 13(20):11408, 2021.

[54] Francesco Zullo, Gianluca Fazio, Bernardino Romano, Alessandro Marucci, and Lorena Fiorini. Effects of urban growth spatial pattern (UGSP) on the land surface temperature (LST): A study in the Po Valley (Italy). *Science of The Total Environment*, 650:1740–1751, 2019. ISSN 0048-9697. doi: https://doi.org/10.1016/j.scitotenv.2018.09.331. URL https://www.sciencedirect.com/science/article/pii/S0048969718337884.