Towards Agenda 2030: Use of GIS in visualizing emissions from personal automobiles for evidence-based policy and planning

Jöran Matson
Towards Agenda 2030: Use of GIS in visualizing emissions from personal automobiles for evidence-based policy and planning

Jöran Matson

Supervisor: Martin Wetterstedt
Subject Reviewer: Peter Lazor
Table of Contents

1. Introduction ................................................................................................................. 1
   1.1. Background ............................................................................................................. 1
   1.1. Research gap and rational of study ................................................................. 2

2. Theoretical Framework ............................................................................................. 4

3. Methods ....................................................................................................................... 4
   3.1. Study Delimitations ............................................................................................... 4
   3.2. GIS Software .......................................................................................................... 5
   3.3. Data .......................................................................................................................... 5
       3.3.1. Security ............................................................................................................ 5
       3.3.2. Pre-Processing ............................................................................................... 5
       3.3.3. Emissions Calculations .................................................................................. 6
       3.3.4. Geocoding ...................................................................................................... 6
       3.3.5. Statistical Analysis ......................................................................................... 7
   3.4. Visualization Methods .............................................................................................. 7
       3.4.1. Choropleth ....................................................................................................... 8
       3.4.2. Heatmap .......................................................................................................... 8
       3.4.3. Spatial Interpolation ....................................................................................... 9
       3.4.4. Hot Spot Analysis ......................................................................................... 12
       3.4.5. Buffer Averaging ............................................................................................ 12
   3.5. Classification Methods and Visual Representation ............................................... 14

4. Results .......................................................................................................................... 15
   4.1. Cleaning and Verifying .......................................................................................... 15
       4.1.1. Pre-processing ............................................................................................... 15
       4.1.2. Statistical Analysis ....................................................................................... 15
       4.1.3. Validating the data ...................................................................................... 19
   4.2. Visualization of data .............................................................................................. 20
       4.2.1. Choropleth .................................................................................................... 20
       4.2.2. Kernel Density ............................................................................................... 21
       4.2.3. Spatial Interpolations ................................................................................. 22
       4.2.4. Hot Spot Analysis ....................................................................................... 24
       4.2.5. Buffer Averaging ......................................................................................... 25

5. Discussion ..................................................................................................................... 26
   5.1. Putting It All Together: A Suite of Insights and Policies ...................................... 26
   5.2. A Deeper Dive into Evidence Based Policy: Developing Public Transportation as an Example ................................................................. 31
   5.3. Possible Areas for Improvement and Recommendations for Further Studies .......... 34
6. Conclusion ........................................................................................................................................36
7. Acknowledgements .........................................................................................................................37
8. References ......................................................................................................................................38

Appendix A. Data Security ..................................................................................................................43
Appendix B. Distribution of data in visualizations due to classification methods .................................44
Appendix C. Buffer Averaging Script Logic .......................................................................................49
Appendix D. Buffer Averaging Python Script ....................................................................................50

Table of figures

Figure 1: Test scenario input and hand calculations – each corner of the square was given a value and a buffer area with the radius equal to the width of the square. The average of these values where they overlap was calculated by hand. ........................................................................................................13

Figure 2: Results from Python script of test scenario – input values of 9, 2, 8, and 5 with their respective positions as shown in Figure 1 were fed into the python script. The script then calculated the average values for where the respective buffers overlap and generated the output seen above. 13

Figure 3: Histogram of emissions data – the y-axis represents the frequency values appear in the dataset and the x-axis represents said value. The range of the top 100 and top 1% of values are indicated to display the low frequency throughout this range. .................................................................16

Figure 4: Histogram of emissions data – outliers removed..................................................................18

Figure 5: Normal QQ Plot of emissions data ........................................................................................19

Figure 6: Comparison of datapoint location (left) with population density (right) ............................20

Figure 7: Average yearly emissions of postal code areas presented using quantile classification – a breakout section of the urban center is provided for clarity due to the small size of postal codes in the urban area. .................................................................................................................................21

Figure 8: Kernel Density, weighted by Yearly Emissions ..................................................................22

Figure 9: Inverse Distance Weighted Interpolation of Yearly Emissions, Quantile Classification ......23

Figure 10: Spline with Tension Interpolation of Yearly Emissions, Quantile Classification ................23

Figure 11: Ordinary Kriging Interpolation of Yearly Emissions, Quantile Classification ......................23

Figure 12: Overlapping areas of the top quartile of all three interpolation results ..............................24

Figure 13: Hot Spot Analysis of Emissions Data ................................................................................24

Figure 14: Buffer Average. Displays emissions from personal automobiles at the vehicle’s registered address. Values are given in CO2/year as calculated by yearly travel multiplied by car emission factor as stated in Transportsyrelsens registry. Emissions are represented as equal sized circles and where emissions overlap, the mean is created. ................................................................25

Figure 15: Largest emitters within Uppsala municipality .......................................................................27

Figure 16: Emissions per Capita for postal code areas using quantile classification ...........................29

Figure 17: Population Density, populace per thousand m² (ibid.)(left), Figure 8 reproduced for ease of comparison (right) .................................................................................................................................30

Figure 18: Kernel Density (Figure 8) divided by Population density (Figure 17, left) – presented using natural break (jenks) classification ........................................................................................................30

Figure 19: Comparison of Hot Spots and bus route coverage ............................................................32

Figure 20: Comparison of Buffer Analysis and bus route coverage ..................................................32

Figure 21: Comparison of Kernel Density automobile emission visualization with the proposed tram route ........................................................................................................................................................................33
Figure 22: Comparison of Buffer Average automobile emission visualization with the proposed tram route. A reduced radius of 500m is used for the buffers in the displayed buffer analysis for increased resolution in this ‘zoomed-in’ map. .................................................................................................. 33

Figure 23: Frequency distribution of data in the Choropleth map of Figure 7 with the quantile classification method employed .................................................................................................................. 44

Figure 24: Frequency distribution of data in the Choropleth map of Figure 16 with the quantile classification method employed .................................................................................................................. 44

Figure 25: Frequency distribution of data in the IDW map as presented in Figure 9 using the quantile classification method. Frequency distribution for the Spline with Tension as displayed in Figure 10 was omitted due to no perceptible difference from the distribution of the IDW method .................................................................................................................................................. 45

Figure 26: Frequency distribution of data in the Ordinary Kriging as shown in Figure 11. A manual classification method was used to match the class ranges of the IDW and Spline Tension methods to ease comparison between the methods .................................................................................................................................................. 46

Figure 27: Frequency distribution of Hot Spot Analysis as shown in Figure 13. ‘-3’, ‘-2’, and ‘-1’ correspond to 99%, 95%, and 90% confidence of cold spots, respectively. The same goes for the hot spots, but these values are positive. .................................................................................................................................................. 46

Figure 28: Frequency distribution of the buffer analysis as shown in Figure 14. Standard deviation is used here as the classification method with an interval size of 1/2 standard deviation .................................................................................................................................................. 47

Figure 29: Frequency distribution of the buffer analysis as shown in Figure 20. Quantile distribution is used here as the classification method .................................................................................................................................................. 47

Figure 30: Frequency distribution of the buffer analysis as shown in Figure 22. Quantile distribution is used here as the classification method .................................................................................................................................................. 48

Table of Tables

Table 1: Summary of Mean Values – emission values are broken down into more basic variables of fuel efficiency and distance driven to investigate why such large emission levels appear in Figure 3 .................................................................................................................................................. 16

Table 2: Comparison of statistical values of distance driven, before and after removal of outliers compared to data on all of Sweden. .................................................................................................................................................. 17

Table 3: Summary of mean values – outliers removed .................................................................................................................................................................................................................................................. 18
Towards Agenda 2030: Use of GIS in visualizing emissions from personal automobiles for evidence-based policy and planning

JÖRAN MATSON


Abstract:
Many countries have committed to goals of emission reductions outlined by the Paris Agreement. The actions and results seen thus far from national governments have been lackluster notwithstanding their good faith pledges. City governments have taken it upon themselves to pick up where their national governments have failed.

A large portion of emissions are due to transportation. Individuals’ daily travel patterns are typically limited to within a relatively local area. As mobility patterns differ between cities, municipalities have a unique role in transitioning to a sustainable society that blanket policies by national governments cannot achieve. This results in a bottom-up approach in achieving national commitments to the Paris Agreement.

In order to make effective policies and plans, municipalities should make decisions based on known information. Before reducing emissions, it must be known where and by whom the emissions are being produced. This report uses Uppsala as a case study to explore how GIS can be used to communicate and create an understanding between data scientists and politicians so that mitigation efforts can be evidence-based. The report results in several methods for visualizing personal automobile emissions based on registration data. The report continues in discussing some potential actions that can be taken to addressing the emissions from neighborhoods indicated as large contributors by the visualizations.

Keywords: Sustainable Development, Geographic Information Systems, GIS, carbon emissions mitigation, local policy, city planning, co-modality, multimodal, mobility, transportation

Jöran Matson, Department of Earth Sciences, Uppsala University, Villavägen 16, SE 752 36 Uppsala, Sweden
Towards Agenda 2030: Use of GIS in visualizing emissions from personal automobiles for evidence-based policy and planning

JÖRAN MATSON


Summary:

Many countries have committed to goals of emission reductions outlined by the Paris Agreement. The actions and results seen thus far from national governments have been lackluster notwithstanding their good faith pledges. City governments have taken it upon themselves to pick up where their national governments have failed.

A large portion of emissions are due to the reliance on automobiles. An understanding of local mobility patterns is fundamental for city government and planners to promote co-modal mobility. This results in a bottom-up approach in achieving the national emission mitigation goals. GIS is used in this study to investigate mobility patterns, specifically in examining which neighborhoods have higher emissions associated to a reliance on automobiles for mobility needs. Methods including statistical analysis along with visualization techniques such as choropleth maps, heatmaps, spatial interpolation, and hot spot analysis are used to analyze the data and reveal existing patterns. Furthermore, a custom technique is developed to visualize the data.

To illustrate how GIS can be used for automobile emissions mitigation within the municipality, visualizations of emissions are overlain with public transportation accessibility. Both Uppsala’s current bus routes are investigated as well as proposed routes for a new tram system. Cycle paths or ridesharing services can be evaluated in a similar fashion.

The report provides a foundation for scientists to communicate and cooperate with local governments to create data-driven policies and planning decisions in promoting co-modal mobility in an effective manner.

Keywords: Sustainable Development, automobile emissions, mobility, GIS, data visualization, Agenda 2030, local policy, city planning, co-modality, multimodal, transportation

Jöran Matson, Department of Earth Sciences, Uppsala University, Villavägen 16, SE 752 36 Uppsala, Sweden
1. Introduction

1.1. Background

Mobility is a central part of our today’s society. Our transportation networks provide access not only to goods and services but also social and economic opportunities. This understanding creates many links between mobility and the three pillars of sustainable development: the economic, societal, and environmental spheres. With issues such as conflicts over oil, urban sprawl, and air quality linking the respective pillars of sustainability to the automobile, it is clear that an overhaul of our auto-centric transportation infrastructure and the culture surrounding it is needed in order to create a more sustainable society (Schwägerl 2019).

While Sweden has a reputation of being an exemplar of creating a sustainable society for the so called developed nations, the rhetoric of Swedish politicians and the actions taken in physically structuring our society do not align (Fenton & Gustafsson 2015). Although all three pillars of sustainable development are important, recent events in the political sphere have brought the environmental pillar into the spotlight in the form of climate change mitigation and adaptation. Due to Swedish voters’ concern of environmental issues steadily growing (von Heijne 2019), it is no surprise that politicians on the national level campaign to dramatically reduce greenhouse gas emissions. Regardless of Sweden’s national goals to follow the Paris Agreement, the reduction of emissions from automobiles has stalled with the levels remaining virtually unchanged in recent years. In fact, the latest national statistics from 2018 show an increase of 0.5% (Joanna Dickinson 2019).

The trend reversal - from a decline in emission back to an increase in emissions - can be explained by examining two main variables: distance driven and fuel efficiency. The fleet of Swedish registered vehicles is continuously being driven an increased number of kilometers; a record high was achieved in 2018 (Trafikanalys 2018a). In the past, advancements in technology (Naturvårdsverket 2016) increasing the fuel efficiency of internal combustion engines (ICEs) at an average rate of 1.5% per year (IEA 2019) and a slow growth of electrified and biofuel powered vehicles to the market share was enough to counter the increase in kilometers driven. However, the slow rates of increases in fuel efficiency and transition to non-fossil fuel powered vehicles is no longer enough to counteract the increase in distances traveled (Joanna Dickinson 2019). Additionally, the average weight of passenger vehicles has tended to increase, counteracting the efficiency increases of ICE technology (Willerström 2019).

Counter to records being set by the Swedish fleet for distance traveled, the average distance traveled by individual automobiles has reached a record low (Trafikanalys 2018a; Jensen 2019). The cause of the contradicting statistics of decreased average distance driven by individuals with an increased distance driven by the Swedish automobile fleet can be attributed to more widespread ownership. Purchasing a new heavy vehicle has been the favored decision rather than electing to carpool or travel by alternative means (Willerström 2019).

Climate scientists have long warned about the dangers of emitting carbon into the atmosphere (McNutt 2015). Scientific studies of the climate have begun to refocus on being applicable to policy makers. For example, a study conducted by NASA studied emissions by sector and concluded that emissions from motor vehicles is the leading cause of climate change (Voiland 2019). The idea of replacing ICEs with electric vehicles (EVs) has become a mainstream development strategy to tackle the challenge of climate change. For this idea to truly decarbonize the transport sector the generation of electricity, which at the current moment is mostly achieved through the use of fossil fuels, would have to be decarbonized as well. Even if the transition to EVs would be achieved, such a technocratic solution maintains and possibly even promotes the expansion of current infrastructure and commuting behaviors that are energy intensive, both in the development and use phases. While it can be argued that the potential to decarbonize the transportation sector lies in a transition to EVs, it entirely neglects the social and economic pillars of
sustainable development. Furthermore, it introduces new environmental problems such as the large
demand of rare earth minerals and the damage that mining for these could cause.

Co-modal mobility – defined as “the efficient use of different modes [of transportation] on their own and
in combination” (European Commission 2006) – provides an approach for decarbonizing transportation.
Rather than relying solely on the automobile for transportation, puzzle pieces that hold potential include
more reliance on cycling; expanding public transportation; reducing the size of automobiles to dimensions
more appropriate for the task; replacing personal ownership of vehicles with a sharing economy;
replacing fuel taxes with pay per use; as well as the automation of vehicles (William Todts 2018).
Research reveals that co-modal transportation not only cuts down on emissions, but also provide social
and economic benefits (Guillermo, Panayotis, Hande & Mert 2013).

The Royal Swedish Academy of Engineering Science, or Kungliga Ingenjörsvetenskapsakademien (IVA)
recently released a report where they state that “In order to reach the climate target in 2045, a
significantly faster pace of change and a radical transformation of the transport system, but also society as
a whole, with an aim of net zero emissions is needed” (Folkesson et al. 2019, p. 28). On the national
level, the prime Minister of Sweden committed to prohibiting the sale of new automobiles powered by
fossil fuels after the year 2030 (Stefan Löfven 2019). Current policy focuses on replacing old vehicles
with new vehicles, even though measures to reduce emissions in the use phase – such as promoting
alternatives to driving with common examples being various forms of public transportation or bicycling –
are more effective and have previously existed (Folkesson et al. 2019). While the national government
continues with the use of status quo strategies to uphold the Paris Agreement, institutions around Sweden
have begun investing large amounts of resources towards understanding how local governments can
courage and support co-modal mobility through better city planning (Energimyndigheten 2016;
Alfheim 2019).

1.1. Research gap and rational of study

Issues concerning transportation are inherently a geographical problem. When discussing emission
reductions, it is common to speak in terms of national goals. While the large scale is important to
understand, individuals’ daily mobility patterns are typically limited to a more local level (González,
Hidalgo & Barabási 2008; Liu, Wang & Ye 2018). Thus, to achieve national emission reduction goals,
regional mobility solutions that consider the economic, social, and cultural context unique to each area are
required (Moavenzadeh & Markow 2007). This echoes the insight of the ‘think globally, act locally’
proverb. With a bottom-up approach in adjusting local infrastructure and behavioral spatial mobility
patterns of individuals, the reduction in national emission statistics can be more effectively achieved.

To understand emissions, many cities commonly have and still use air quality as their metric. This is an
established method that has been developed over multiple centuries. Consequently, there exists
considerable legislation setting requirements for air quality. While these measurements provide detailed
information about what the air is composed of, air quality measuring stations are typically located
exclusively in urban areas and give few measurement locations for each city (Kuhlbusch et al. 2013).
Such information is useful in creating general understanding of the situation, however, is not very helpful
in creating an understanding of mobility patterns and needs. Actions, especially legislative, have long
been based on air quality data, but understanding emissions and mobility requirements on a more
individual level could result in more effective mitigation methods and policies. Although technological
developments may possibly provide a new paradigm in air quality measurements that provide deeper
understanding, they are not currently available and need time to develop (Snyder et al. 2013).

---

1 Translated to English by the author from the original version published in Swedish.
What is currently available is information regarding individuals’ driving behavior. Studies such as the one conducted by Perumal and Timmons (2017) use survey data to understand how commuting behavior varies between urban and rural areas. Other studies such as one by Wang and Chen (2015) go one step further by seeking to understand socio-political factors in understanding the accessibility of a community. Both these studies focus on the USA where only samples of the data are available, or data needs to be interpolated. Sweden has an advantage in that it has a history of keeping a rich and detailed registry of its residents. A study by Blind et al (2018) uses data from this registry in its goal of understanding how individuals commute in Sweden. Although Blind et al provide a successful beginning in understanding how a wide variety of parameters correlate with driving behavior, this study will approach the issue from a slightly different angle.

Instead of understanding all modes of transport utilized in daily commuting patterns, this study limits the scope by examining the spatial characteristics of automobile reliance with an aim to identify neighborhoods where this behavior can easily be minimized. This will be done by importing data from Sweden’s automobile registration database into Geographical Information Systems (GIS) software where it can be processed and analyzed.

GIS has been defined many times, but in essence is “a decision support system involving the integration of [spatially] referenced data in a problem solving environment” (Peuquet & Marble 2003, p. 62) where computer software is used to assist in capturing, storing, retrieving, manipulating, and visualizing the spatial data (Clarke 1986). Advances in technology continue to make information more accessible, which in turn enables and encourages planning policy to be evidence based (Wong, Baker, Webb, Hincks & Schulze-Baing 2015). Ronald Abler of the National Science Foundation expressed the importance of GIS with the analogy that "GIS technology is to geographical analysis what the microscope, the telescope, and computers have been to other sciences" (Peuquet & Marble 2003, p. 63). Combining database management abilities with map visualization is the fundamental reason GIS is such a useful technology (Nyerges & Jankowski 2009, p. 41).

This study will use Uppsala municipality as a case study for exploring driving habits of residential areas to understand how the CO₂ emissions from automobile usage are distributed between households. The geographical scope of Uppsala was chosen due to its coherence to the Swedish trend of increased vehicle registration despite co-modal transportation being readily available. Uppsala municipality has received much attention recently for its environmentally progressive developments. For example, it retained its title as Sweden’s most cycle friendly city (Forslund 2019) and won the World Wide Fund for Nature’s 2018 One Planet City Challenge for its high ambitions for eliminating carbon emissions by 2030 (WWF 2018, 2019). Such distinctions and ambitions provide an interesting backdrop to study the resilience of automobiles and the culture surrounding them.

For the Paris Agreement to be upheld, each nation needs to do their part. For Sweden to achieve its goals, actions need to be taken on a local level by each and every municipality. When environmental policies are considered, three common categories of questions are asked: 1) Who is affected, how are they affected, and to what magnitude are the affects? 2) Is it a policy problem? 3) What can be done, will it work, and what are the risks/costs? (Tomich et al. 2004).

Evidence based policy is key in creating effective and lasting environmental policies. The discovery and understanding of evidence are often left to scientists while policy creation is done by the politicians. For the two domains to collaborate, it is crucial for scientists to clearly communicate their findings. Scientists commonly use technical language inaccessible to the public and politicians, thus failing to communicate (Likens 2010). To provide a foundation for building an answer to the above questions in regard to automobile emissions, this report aims to provide a suite of graphs and images that aid scientists to effectively communicate to politicians which geographical areas could be targeted by policies to reduce automobile emissions.
Research Question: Can visualizing the origins of automobile emissions with the use of GIS provide a foundation for evidence-based decision making to increase co-modal transportation?

2. Theoretical Framework
Interpreting and creating an understanding from hundreds of thousand rows of data in text form is near impossible for the human mind. Statistics is one instrument used to aid the understanding of large amounts of data by truncating it into a more digestible form. However, statistics alone is not the most efficient method for conveying geographical information. Geographical Information Systems (GIS) is a tool that has been developed to process geographical data and create graphical visualizations of it to help solve geospatial problems (Lloyd 2010).

In order to understand automobile usage within Uppsala, the data must be investigated. As GIS is a general tool used to analyze specific data, many different techniques have been developed to translate data into visualizations. Each technique presents a slightly different perspective of the story contained within the data. This study will be exploratory in nature, employing an assortment of techniques to fully understand the story told by the dataset.

This study was also used as an opportunity for the author to explore the capabilities of GIS and gain further competencies using this tool. In short, the exploratory nature of this study refers to the exploration of both the data and the capabilities of the software.

3. Methods
This section discusses the obtaining of data and its handling, visualization techniques employed, and the limits of the chosen approach.

3.1. Study Delimitations
Every vehicle that utilizes public roads is required to register and undergo a yearly inspection. However, this study is not concerned with commercial activity, but commuting patterns. Thus, only private vehicles will be considered in this study. This delimitation applies to private vehicles used for commercial purposes as well. A common example of this occurrence can be seen in platforms such as Uber that enable individuals use their own resources to provide a service. The scope excludes these as the study is interested in the use of personal vehicles for commuting. While the taxi may be privately owned and the registration data accurate, this report considers taxis as a means of ridesharing. Research shows that while taxis employ traditional ICE vehicles, allowing it to be in the mix of co-modal transportation does not negatively impact other modes, but promotes the driving segment to exchange their personal vehicle for using a combination of public transportation and taxis (Lee, Jin, Animesh & Ramaprasad 2018).

Another limitation is the quality of the data. Since Transportstyrelsen is the only organization that collects automobile registration data there is no secondary source that can verify the completeness and quality of the dataset. With law mandating individuals to register their automobile and have it inspected yearly, concerns regarding the completeness of the dataset can be dissolved. Regarding the quality of the dataset, information of each individual vehicle is collected by inspection technicians who are required to certified
through competency standards set by transportstyrelsen (Transportstyrelsen 2015). While this should ensure the quality of data collected is consistent, human errors cannot be entirely ruled out.

Lastly, the scope of this study is limited to using various tools and methods the chosen software is capable of in creating an understanding of the emissions data from personal automobiles. The chosen methods are explained in greater detail in Section 3.4. Other software may provide methods not considered in this report.

3.2. GIS Software
The GIS software chosen for use in this study is the ArcGIS suite by the Environmental Systems Research Institute (ESRI). As this is the prominent GIS suite taught at Uppsala University, accessibility to the software and technical support from both the software developer and knowledge sources within the university were driving factors in the decision. Two programs were used from the ArcGIS suite: ArcGIS Online (AGOL) and ArcMap. ArcGIS Online (AGOL) is a web browser-based GIS program within ESRI’s suite that offers geocoding services using their servers (See Section 3.3.4). While AGOL enables geocoding with little difficulty, its functions for managing and manipulating data are limited. ArcMap is a stand-alone desktop program that provides a broader range of much more advanced tools to process and analyze data with. Version 10.6.1 of ArcMap is used for this study as this was the latest version available through the university.

3.3. Data
Automobile Registration data was obtained from Transportstyrelsen, which is the governmental agency responsible for regulating and inspecting all transportation systems in Sweden. The file received included 106,999 data points including information regarding the address of the registered vehicle, vehicle type and year, date registered, inspection date, odometer reading, previous inspection date, previous odometer reading, European emissions standard of the registered vehicle, and fuel efficiency. The data was received in a delimited flat file format and imported into Microsoft Excel for ease of use.

3.3.1. Security
Granted that anyone with access to data containing private information should be mindful of ethical implications and potential consequences of handling such information irresponsibly, the scandal surrounding Transportstyrelsen and their careless outsourcing of data processing (Ohlin 2018) gives reason to be particularly careful with the data used in this study. Actions and guidelines regarding how the data will be handles during this study can be found in Appendix A.

3.3.2. Pre-Processing
Just like any other data, prior to processing and analysis it needs to be ‘cleaned’. The cleaning process removes data points that do not provide value to the study. This may be for reasons such as the data points are outside the scope of study, contain erroneous data, or simply contain no data at all regarding the variables being studied. This both reduces the time needed to process the data and increases the accuracy of the results.
One filtering process is geographical. The data must first have a valid geographical location as well as be within the field of study. This process is described in more detail in Section 3.3.4.

Once the data is confined to a geographical location, attention is turned to the information the data set contains. This study is interested in automobile emissions. Data obtained from Transportstyrelsen provides three quantities that can be used to calculate this: current odometer reading (mätarställning), previous odometer reading (föregående mätarställning), and CO₂ emissions (CO₂ utsläpp). See Section 3.3.3 for an explanation of how these values are used in calculating emissions. All three values are needed to calculate the emissions per annum for a data point; thus, data points will be excluded from the study if any one of the values are missing.

Because the registration dates are not precisely a year after the previous registration, Equation 1 below normalizes the period between measurements to exactly a year. For some data points, the time span between measurements is a few weeks to a few months. Such short intervals may be due to vehicle issues and driving within such short periods may not be representative. Three months was chosen as the cut-off value for including data points in the analysis.

3.3.3. Emissions Calculations
The automobile registration data obtained from Transportstyrelsen provides sufficient data to calculate the yearly emissions of each vehicle and tie these emissions to a geographical location. The equation to calculate the emissions is as follows:

Equation 1: Calculation of Emissions

\[ \text{Emissions per annum} = \frac{365}{t_c - t_i} (d_c - d_i) E_m \]

Where:
- \( d_c \) is the current (at time of registration) odometer reading in kilometers,
- \( d_i \) is the ‘initial’ odometer reading from the previous time of registration in kilometers and
- \( t_c, t_i \) is the time difference in days between the odometer readings
- \( E_m \) is the CO₂ emissions rate of the vehicle in grams per kilometer

Due to registration data not providing enough information to discern whether distances were driven in urban or rural areas, the value for ‘mixed driving’ is used for \( E_m \) in all instances.

3.3.4. Geocoding
In order to create visualizations of the data, each data points must first be located relative to the other data points. Many programs exist that can process information in, for example, the commonly used Cartesian coordinate system. However, the registration data used in this study was given with location information in the form of street addresses and postal codes. The process of converting location codes - such as the form the automobile registration data is given in - into other coordinate systems is known as geocoding (Mayhew 2015).

Based on reasoning presented in Section 3.3.1, the dataset was reduced prior to being uploaded to ESRI’s servers. Prior to geocoding, each data point was reduced to contain only three data fields: the street address, postal code, and city. Each data point was also given a unique identification number in order to concatenate the removed data together again when the online portion of the workflow finished.
In order to import the data set into AGOL, it was converted into a CSV (Comma Separated Value) file. Once imported, the geocoding script within AGOL was run and the results were returned. The results could then be individually reviewed and confirmed as correct or updated with a new location. For some data points, no match was found. In these cases, location information can be inputted manually or removed from the data set. Once the geocoding of the data was reviewed, it could be exported, associated with the emissions calculated in the previous section, and imported into ArcMap for analysis.

Once the datapoints have been located, the distribution of the dataset can be compared to population density. Significant deviations between the two likely is an indicator of errors in the dataset.

### 3.3.5. Statistical Analysis

Similar to how the geographical distribution of the dataset can function as a check to verify the dataset, a statistical analysis of the values being studied can also function as a check for errors in the dataset. Techniques such as histograms, quantile-quantile (QQ) plots, and semi-variogram/covariance clouds can give insight to the distribution of the dataset, if and/or how many outliers are present, how the datapoints relate to each other, and the identification of trends. As no secondary source is available to compare the data with, applying rational judgement against the results of these techniques can function as a validity check of the dataset.

Histograms provide a graphical summary of the frequency distribution of the data in bar chart form. The range the data covers is classified into ‘bins’ that are represented by each bar. The number of data points in each bin are represented by the height of the bar. Such a visual representation enables quickly creating a basic understanding of the spread of the data and the nature of its standard deviation as well as the shape of the distribution.

QQ plots also provide a method for comparing the frequency distribution. Unlike histograms, this is done by plotting attributes of two different data sets against each other in a scatter plot. A normal QQ plot is a specific type where the dataset of interest is plotted against a gaussian, or standard normal, distribution. Comparing data to a gaussian distribution allows for understanding of the skewness of the data and identifies the outliers.

Tobler’s first law of geography states that “near things are related to everything else, but near things are more related than distance things” (Miller 2004). Variograms provide a visual for understanding the extent that this law holds for the given data. Plotting the semi-variance of the dataset paints a picture from which the spatial autocorrelation of the dataset can be quickly understood. Semi-variograms, also known as covariance clouds, plot the covariance for all pairs of data points against the distance between the two points.

### 3.4. Visualization Methods

This section outlines and explains the visualization methods that were explored in this study for analyzing and creating an understanding of emissions data. As previously discussed, each one provides a unique insight into the studied dataset. Some visualization techniques were chosen due to their popular use, others were selected to fill a void left by the shortcomings of other techniques, and yet others for the unique perspectives they provide.
3.4.1. Choropleth

Choropleth maps are the most common method for studying and illustrating socioeconomic data, especially that which is concerned with rates or percentages. To understand and present patterns in the data, the study area is divided up into smaller sub areas. These boundaries can either be naturally occurring (for example dividing the area by land cover type) or imaginary boundaries (such as national borders). The subareas are then colored or shaded corresponding to attributes of the data being studied (Kemp 2007).

A standardized approach of dividing Sweden’s counties into smaller areas for specifically studying emissions data does not currently exist; however, postal codes provide a standardized partitioning method that is commonly used for various purposes. Among many other factors, postal codes are based on population density. The urban/rural divide is commonly a focus in studies of automobile emissions. Due to a lack of an existing scheme for partitioning the municipality that is more pertinent, postal codes will be used as they provide a convenient method that suffices for this study.

3.4.2. Heatmap

Heatmaps, also known as density function visualizations, is a tool that aids in identifying areas where data or data attributes are more concentrated. ArcMap provides a few different methods for producing such visuals, but Kernel Density Estimation is the most relevant for this study.

The basic idea of a Kernel Density Estimation is to construct a smooth curved surface, represented as a continuous color distribution, from a set of discrete data points. The volume under the surface represents the intensity of the data. For example, areas of the surface with a higher elevation, indicated by colors on the high end of the spectrum, represent a higher density or value of the data points.

The most common technique of creating a heat map is the kernel density estimation method (Perrot, Bourqui, Hanusse, Lalanne & Auber 2015). This is slightly similar to a choropleth map in that bounded areas are created and colored. These bounded areas are quite different in their composure. A kernel density estimation produces a raster output, which is made up of ‘cells.’ Contrary to Choropleth maps using large boundaries, the cells of the raster are relatively small and in the form of a homogeneous square lattice (ESRI 2019a). While the squares are colored, they are much smaller than the boundaries used in choropleth maps and provide a gradient rather than abrupt changes in color at boundaries. The default area of each raster cell is calculated as shown by Equation 2. The cell size can be manually changed to either a smaller value to increase the resolution of the output or a larger size for more prompt results.

Equation 2: Default Raster Cell Size

\[
\text{Default raster cell size} = \frac{\min(Width, Height)}{250}
\]

As each cell is a valuation based on surrounding data points, more complex calculations are required in producing this visualization than for choropleth maps. The basic concept is: first a smooth curved surface is calculated for each data point where the peak is at the location of the data point, reducing to zero at a radial distance from said location. This surface is produced by the kernel function. The peak value can be ‘weighted’ by an attribute of the data set. For example, if studying population density, a single data point representing a household may have a ‘population’ value of 3 representing a single child family. The value of three would then be represented in the kernel surface. In this way one data point can be counted

\(^2\) Indicates that the smaller of the two options is used.
multiple times. The value of each raster cell in the output surface is then calculated by summing the values of all kernel surfaces that overlap the said raster cell. To see exactly how this is calculated in ArcMap, refer to the formulas given below.

**Equation 3: Kernel Density Estimation Function (Sucharita Ghosh 2018)**

\[
f(x) = \frac{1}{nb} \sum_{i=1}^{n} K \left( \frac{x - x_i}{b} \right)
\]

Where:

- \( K \) is the value of the kernel function (see Equation 4)
- \( h \) is the search radius (see Equation 5)

**Equation 4: Quartic Kernel Function\(^4\) (Silverman 1986)**

\[
K(x) = \begin{cases} 
3\pi^{-1}(1 - x^T x)^2 & \text{if } x^T x > 1 \\
0 & \text{otherwise}
\end{cases}
\]

**Equation 5: Search Radius for Kernel Density Estimation\(^5\) (ESRI 2019a)**

\[
b = 0.9 \times \min\left( SD, \frac{1}{\sqrt{\ln(2)}} \times D_m \right) \times n^{-0.2}
\]

Where:

- \( SD \) is the standard distance
- \( D_m \) is the median distance
- \( n \) is the number of points or – if a population field is provided – the sum of the population field values.

The search radius functions as a smoothing parameter. A larger bandwidth can potentially ‘over-smooth’ the results while a smaller number may allow for unwanted spikes in the results. Equation 5 above is based on the normal distribution approximation developed by Silverman but modified to minimize the ‘ring around the points’ anomaly that is often observed when working with sparse datasets (Silverman 1986; ESRI 2019a).

### 3.4.3. Spatial Interpolation

Another method of analyzing geospatial point data is by fitting a raster surface to the known points. This is a useful technique when known data is not geographically dense as known data points can be used to estimate unknown values. Numerous interpolation methods exist, but almost all interpolation methods follow the same basic logic as expressed in Equation 6 with small variations (Li & Heap 2014).

---

3 The formula actually used in ArcMap may differ from the one given here. The only information provided by ESRI is that their function is “based on” this quartic kernel function.
4 Kernel function evaluation \( k(x_i, x_j) \) equivalent to \( x^T x \) is used in kernel methods (Martin 2005)
5 The search radius is also occasionally referred to as the ‘bandwidth’
6 Indicates that the smaller of the two options is to be used.
Equation 6: General Interpolation formula

\[ Z(s_0) = \sum_{i=1}^{n} \lambda_i Z(s_i) \]

Where:
- \( Z(s_i) \) is the measured value at the /th location
- \( \lambda_i \) is the (unknown) weight for the measured value at the /th location
- \( s_0 \) is the measured location and
- \( n \) is the total number of measured locations.

Inverse Distance Weighting

The first interpolation method explored is the Inverse Distance Weighted (IDW) technique, where an unknown value is calculated using surrounding known values as shown in Equation 6.

IDW is a very self-describing method as it does just that; it assumes that as the distance between two features increases, the less similar the values are. The weight, \( \lambda \), that a known value has on an unknown value is given by Equation 7.

Equation 7: Inverse Distance Weighting function

\[ \lambda_i = \frac{1}{d^P} \]

Where:
- \( d \) is the distance between the known value to the unknown value and
- \( P \) is a power chosen by the cartographer

ArcMap uses a default value for \( P \) of 2. A smaller number gives more weight to features further away while a larger number gives more weight to features further away. An ‘optimal’ power value can be calculated with the root mean square prediction error (RMSPE) which is a statistic that quantifies the error of the prediction surface to the known values. The optimal power value is one that has the smallest RMSPE.

As values far from the estimate being calculated have little influence on the result, they can be excluded by creating search neighborhoods. Only using values within the neighborhoods to calculate the estimate hastens the calculation. ArcMap allows for adjustment of the shape of the neighborhood as well as adjusting the minimum and maximum values used in calculating the estimate.

Spline

Similar to IDW, a spline interpolation is also an exact interpolation, meaning that the output raster surface passes through all the input data points. Where they differ is that the spline technique is not bounded by the values of the known points meaning that values can be predicted that are more extreme than the known values. ArcMap provides two types of spline interpolation methods; regularized spline and tension spline. Of the two, tension spline is more suited for handling more drastic variances in data and thus will be the one included in this study.

A tension spline, similar to IDW, fits a mathematical formula to the nearest points while also intersecting known data points and minimizing surface curvature. The formula used in calculating the spline raster surface varies slightly from the one used in the IDW method and is provided in Equation 8.
Equation 8: Tension spline formula

\[ S(x, y) = a + \sum_{i=1}^{n} \lambda_i R(r_i) \]

Where:
- \( a \) and \( \lambda \) are coefficients found by the solution of a system of linear equations
- \( r \) is the distance from the point \((x, y)\) to the \(i\)th point
- \( R(r) = \frac{1}{2\pi\varphi^2} \left[ \ln \frac{r\varphi}{2} + c + K_0(r\varphi) \right] \)

\( \varphi \) is the weighting parameter
- \( K_0 \) is the modified Bessel function
- \( c \) is a constant of 0.577215

The weighting parameter is a non-negative value chosen by the cartographer. A larger value results in a more coarse output surface.

Kriging

The IDW and Spline interpolation methods are both deterministic functions as the results are directly dependent on surrounding known values and mathematical formulas that determine the smoothness of the interpolation. The Kriging method differs from deterministic functions as it is a geostatistical function, meaning that it depends on statistical relationships between the known points to calculate interpolation. Although not deterministic, the Kriging method is quite similar to the IDW method as they both use Equation 6 and output a raster surface. Where the methods differ is in how the weighting is calculated; IDW looks only at the distance between values while Kriging goes one step further and considers quantified special autocorrelation. In short, the weight \( \lambda_i \) is dependent on a fitted model to the known points, distance to the prediction location from the known points, as well as the spatial relationship between the known values surrounding said prediction point.

Equation 9 gives insight into the assumptions made by various kriging methods to create a prediction surface and a corresponding error map.

Equation 9: The Kriging Model

\[ Z(s) = \mu(s) + \varepsilon(s) \]

Where:
- \( s \) represents the location
- \( Z(s) \) is the variable of interest
- \( \mu(s) \) is a truncated trend of the variable of interest
- \( \varepsilon(s) \) is the error between \( \mu \) and \( Z \) for each \( s \).

A few variations of kriging exist, each made more suitable for different types of input data by using slightly different assumptions regarding \( \mu \). Simple kriging assumes \( \mu \) is a known constant; ordinary kriging assumes \( \mu \) is an unknown constant; and universal kriging assumes \( \mu(s) \) is a deterministic function. Prior to deciding on which kriging algorithm to use, a geostatistical examination of the data should be done to understand which assumptions are most applicable. Section 3.3.5 outlines relevant tools for this examination.
3.4.4. Hot Spot Analysis

Hot spot maps and heatmaps may sound as if they are equivalent or similar concepts but are calculated using quite dissimilar techniques. While a heatmap function outputs a map conveying simply the magnitude or density of the feature, a hot spot function goes one step further by outputting a map conveying statistical relevance of the data points. In other words, a heatmap is a more sophisticated choropleth map where a hot spot map is a statistical test for randomness.

Statistical testing of randomness can be important as the human mind has a strong capability of finding patterns, sometimes where they do not exist. This can quickly lead to the minds natural tendency to link the pattern with a cause, whether or the pattern exists or is simply random. Once a cause has been established by the human mind, confirmation biased can easily set in (Cairo 2016). Hot spot analysis can prevent falsely identified patterns by reducing the subjectivity by allowing statistical methods to identify the patterns.

A mathematical process for determining if there is a pattern is the Getis-Ord Gi* method. First a null hypothesis, or an assumption that the dataset is completely randomized, is made. Equation 10 is then used to analyze each data point (or an attribute of it) relative to neighboring values. The calculations result in a z-score and p-score that indicate when the null hypothesis can be rejected. P-scores is a measure of how probable the value would appear in a normal distribution and z-score represents the standard deviation of a value within the same normal distribution. The level of the z-score and p-score must reach is based on the confidence level, or degree of risk, that the cartographer is subjectively willing to accept. Confidence levels of 90%, 95%, and 99% are commonly used. When the hypothesis is rejected, statistically significant areas where high and low values are clustered are signified in the map.

Equation 10: Gi* local statistic

\[ G_i^* = \frac{\sum_{j=1}^{n} w_{i,j}x_i - \bar{X} \sum_{j=1}^{n} w_{i,j}}{S \sqrt{\frac{n \sum_{j=1}^{n} w_{i,j}^2 - \left( \sum_{j=1}^{n} w_{i,j} \right)^2}{n-1}} \}

Where:
- \( x_i \) is the attribute value for feature \( j \)
- \( w_{i,j} \) is the spatial weight between feature \( i \) and \( j \)
- \( n \) is the total number of features, and
- \( \bar{X} \) and \( S \) are defined by the following equations:

\[ \bar{X} = \frac{\sum_{j=1}^{n} x_j}{n} \]
\[ S = \sqrt{\frac{\sum_{j=1}^{n} x_j^2}{n} - \bar{X}^2} \]

3.4.5. Buffer Averaging

A custom script was written in python for this study to create a visualization that was missing from the ArcMap toolboxes. The script requires a buffer shape file as an input. This assigns a geographic area to each individual data point. The script then separates each buffer into its own file and converts each one into a raster. The average is then calculated for each instance where these raster files overlap. These averaged values are then compiled into a single file and that the python script outputs as a TIF file. For a
more detailed explanation of the logic, see Appendix C for a flow chart of the logic. The script is provided in Appendix D in its entirety.

In order to verify the computations of the script, a simple test scenario was run. 4 data points with unique values were each placed in the corner of a square. A buffer with a radius equal to the length of a side of the square was given to each data point. This simple dataset was then fed into the script. The results can be seen in Figure 2. Comparing the results of the script with the hand calculations of Figure 1 confirms the script is performing as desired.

**Figure 1: Test scenario input and hand calculations –** each corner of the square was given a value and a buffer area with the radius equal to the width of the square. The average of these values where they overlap was calculated by hand.

**Figure 2: Results from Python script of test scenario –** input values of 9, 2, 8, and 5 with their respective positions as shown in Figure 1 were fed into the python script. The script then calculated the average values for where the respective buffers overlap and generated the output seen above.
3.5. Classification Methods and Visual Representation

Regardless of which method is chosen to create a visualization, how the results are presented has a large impact on how the reader interprets the data. The number of classes, how the classes are sized, and which colors are chosen to represent said classes all influence how the map is understood.

The number of classes used in the map will change depending on the purpose of the map. When using a discrete color code, the choice in number of categories or bins is critical. If too many bins are used, then it is difficult for the reader to distinguish between the shades; too few and the map may lose effectiveness in communicating the data. Research suggests that the human mind cannot distinguishing between groups of 10 or more shades of a color in a single map (Kemp 2007). If it is important to be able to decipher which areas have what values, the number of classes should be 9 or less. If the purpose of the map is only to convey a trend within an area on understanding how values relate to each other is more important than understanding the actual values of each class, then a more continuous (larger number of classes) may be more appropriate. The majority of visuals presented in this study will use the later strategy as the interest is in trends rather than actual values.

The choice of colors is important in creating an understandable map. Using an unreasonable number of hues in communicating a single variable can produce confusion or uncertainty. One hue is the most comprehensible technique; While more colors may be added to increase contrast, using too many colors can lead to a ‘rainbow color map’ that misleads the interpretation through non-data dependent gradients (Borland & Li 2007).

Lastly, how the classifications are created can emphasize different parts of the data. Different classification methods may be used for different situations depending on how the data is distributed or if a certain range of the data set is the focus and it is desired to convey it more precisely. With a dataset that is evenly distributed over the entire range, even intervals can be used. For situations where the dataset is not distributed evenly or a focus on a specific range is desired, methods such as quantile, natural breaks (Jenks), or standard deviation are commonly used. The quantile method is a classification method that breaks the dataset up into a chosen number of ranges with an equal number of datapoints. This results in classes spanning a smaller range of values where the data is more dense and larger ranges where the data is less dense. The natural breaks (Jenks) method constructs classes by identifying ranges where data values are concentrated, and places class breaks between these grouped collections of values. Standard deviation, as the name indicates, creates classes based on the standard deviation which measures the amount each value deviates from the mean of the dataset. The classes are created for either a whole, half, third, or quarter of a standard deviation. ArcMap also provides a method named geometric interval, which combines elements of the quantile and natural breaks (Jenks) methods (ESRI 2019b). An understanding of the data through statistical analysis (see Section 3.3.5) can be helpful in deciding which classification method to use.
4. **Results**
This section presents the outcomes of the methods previously outlined applied to the case study of Uppsala municipality.

4.1. **Cleaning and Verifying**

4.1.1. **Pre-processing**
The first step was to geographically filter the data points. From the 102103 data points that were identified as having a postal code within Uppsala municipality, 1795 data points did not receive a match during the geocoding process. A small portion of these, on the scale of a few hundred, were due to a lag between the recent construction of residential buildings and their corresponding addresses being cataloged in the address database used by AGOL. The majority of unmatched data points were due to an association with a PO box number rather than a physical address. After being geographically filtered the dataset contained 100,308 valid data points.

Of the data points successfully geocoded and located within Uppsala municipality, 17636 data points were missing data regarding the current odometer reading; 29363 did not have a record of their previous odometer reading; and 23803 had no data regarding CO₂ emissions. Note that some data points were missing data in more than one of these areas. Removing the points lacking information brought the usable data points to 50732. Finally, removing data points with a time span of less than a quarter of a year resulted in 40128 usable data points. Figure 6 provides a map where the valid data points are geographically located.

4.1.2. **Statistical Analysis**
A histogram of the cleaned data is provided below in Figure 3. It can be quickly noticed that the spread is quite large, with a maximum of 217 Tonnes CO₂ per year. Such large amounts of emissions by a personal vehicle are remarkably high.
To understand the large right tail (or positive skewness) of Figure 3, factors driving emissions were investigated. While factors such as vehicle weight and age of vehicle are commonly studied, these all effect and can be understood together by analyzing the fuel efficiency; provided fuel efficiency numbers are correct, c.f. Willerström 2019. Thus, fuel efficiency along with distance driven were examined. Table 1 provides a comparison of these values from different portions of the dataset. It can be noticed that fuel efficiency remains relatively similar for each group while a drastic difference can be seen in the distance driven per year for each group.

Table 1: Summary of Mean Values – emission values are broken down into more basic variables of fuel efficiency and distance driven to investigate why such large emission levels appear in Figure 3

<table>
<thead>
<tr>
<th></th>
<th>Mean Fuel Efficiency</th>
<th>Mean Distance Driven</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(g CO₂ / km)</td>
<td>(km / year)</td>
</tr>
<tr>
<td>Entire Dataset</td>
<td>169</td>
<td>15,844</td>
</tr>
<tr>
<td>Top 1%</td>
<td>188</td>
<td>115,330</td>
</tr>
<tr>
<td>Top 100</td>
<td>184</td>
<td>268,190</td>
</tr>
</tbody>
</table>

The large number of kilometers for both the top 100 and top 1% are both quite remarkable. With the largest emitter driving a distance just shy of 900 thousand kilometers within a single year, the vehicle would have to spend 80% of the year, day and night, traveling at 120 km/h nonstop on a major highway. While this is one of the extreme values, such large values seem unrealistic for a personal vehicle. One situation where this amount of driving may be feasible could be taxis shuttling between Uppsala C and Arlanda airport. Based on google maps, this drive is 36.3 km and takes 28 minutes. The maximum distance driven if this route were to be driven non-stop for an entire year is 682 thousand kilometers, falling short of the largest values in the dataset. To achieve the mean for the top 100 and top 1%, it would
require this trip to be driven non-stop for 9.5 and 4.1 hours a day for the entire year. This seems much more reasonable. As discussed in Section 3.1, taxis are outside the scope of this study.

Table 1 brought to light that the data set included values of yearly distance driven that were larger than expected. To determine which points should be considered outliers, statistical values were compared to the latest data for all of Sweden that Transportstyrelsen had collected. Table 2 below summarizes this comparison and reveals a stark difference in magnitude of both the max and mean values. The maximum value of Transportstyrelsen’s dataset, 49266 kilometers, was used to determine the outliers. Data points associated with values larger than this were removed from this study’s dataset. This reduced the dataset by 297 to a total of 39,893 data points. The maximum and mean distance driven were recalculated and presented in Table 2. Notice how the maximum matches, but the mean remains relatively close to the original value. This raises interest in studying the frequency distribution again. A histogram in the same form of that presented in Figure 3 is reproduced below in Figure 4 with the outliers omitted.

<table>
<thead>
<tr>
<th></th>
<th>Maximum Distance Driven (thousand km / year)</th>
<th>Mean Distance Driven (thousand km / year)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Entire Sweden</td>
<td>49.38</td>
<td>1.2</td>
</tr>
<tr>
<td>Dataset including outliers</td>
<td>924.7</td>
<td>15.8</td>
</tr>
<tr>
<td>Dataset after removal of outliers</td>
<td>49.2</td>
<td>12.5</td>
</tr>
</tbody>
</table>

Table 2: Comparison of statistical values of distance driven, before and after removal of outliers compared to data on all of Sweden.

7 Values obtained via email communication with Transportstyrelsen. These values are unpublished at the time of this writing. The mean value was similar to the officially published values of previous years.

8 Data points associated with values of kilometers driven per year larger than 49,266 were considered outliers and removed from this study.
Figure 4: Histogram of emissions data – outliers removed

Figure 4 is much closer to a normal distribution than before and seems more reasonable. Table 3 provides the values presented previously in Table 1, but recalculated with the outliers omitted.

Table 3: Summary of mean values – outliers removed.

<table>
<thead>
<tr>
<th></th>
<th>Mean Fuel Efficiency (g CO₂ / km)</th>
<th>Mean Distance Driven (km / year)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Entire Dataset</strong></td>
<td>169.60</td>
<td>12,508</td>
</tr>
<tr>
<td><strong>Top 10%</strong></td>
<td>196.82</td>
<td>26,511</td>
</tr>
<tr>
<td><strong>Top 1%</strong></td>
<td>217.07</td>
<td>36,257</td>
</tr>
</tbody>
</table>

A normal QQ plot of the cleaned emissions data is provided below in Figure 5. The “U” shape of the graph indicates that the data has positive skew, or that the distribution of data points is towards the left side of the curve. The plot also features heavy tails, indicating that the data is distributed towards the extremes rather than towards the center of a normal distribution (Ford 2015). This tells us that the quantiles increase slower for the smaller values. The larger the values become, the quicker the quantiles increase.
It was not possible to produce a semivariogram/covariance cloud for this dataset due to the software returning an ‘unspecified error’. The following section will look at various visualizations to examine if there exists a correlation between the frequency distribution and any geographical patterns or trends. The filtered dataset with outliers removed as presented in Figure 4 will be used in analysis presented in the following sections.

To summarize insights gained from statistical analysis, the histogram and QQ-plot provided in Figure 4 and Figure 5 respectively show that the dataset is light tailed, also described as positive skew, meaning the distribution has a higher concentration towards the low end than a gaussian (normal) distribution would predict. This causes the high emitters to be responsible for an expressly unproportionally amount. To put this in perspective, the top 1% and 10% are responsible for 3.6% and 23.6% of the total automobile emissions of Uppsala county, respectively. While both fuel inefficiency and distance driven are both positively correlated with emissions, distance driven appears to be a larger contributor, as can be seen in Table 3.

4.1.3. Validating the data
Prior to performing any analysis, the dataset should be validated. Because only one agency collects this data, it is impossible to verify it against another source. Hence, rational judgement is used.

To validate the geographical distribution, the location of the datapoints are compared with population density. These two should not differ drastically. This comparison is provided in Figure 6 where a likeness can be observed.
With the geographical attributes of the dataset confirmed, focus is turned to the values. Comparing the vehicles per capita

Based on the dataset, Uppsala municipality has a mean emissions rate of 2.09 tonnes CO$_2$ from automobiles per year. This roughly matches Sweden’s total emissions from personal automobiles divided by the number of vehicles registered nationwide (Statistiska Centralbyrån 2019a, 2019b). For comparison, the average resident of the United States emits 4.6 tonnes CO$_2$ per year from personal automobile usage (US EPA 2016). This matches general perceptions as well as the author’s personal experience of automobile culture of the two countries.

### 4.2. Visualization of data

#### 4.2.1. Choropleth

The quantile classification method, as explained in Section 3.5, was chosen for the choropleth map (presented below in Figure 7) due to the values being concentrated in a relatively small section of the range. Figure 23 in Appendix B demonstrates how this method provides more distinction between data points where they are more concentrated, around the median (2.02 for these areas, which matches closely with that presented in Section 4.1.2), while grouping the more extreme values together.
4.2.2. Kernel Density

Emissions data represented using a kernel density function is provided below in Figure 8. Each raster cell has a size of 100 square meters. To give equal representation to values throughout the spectrum, a quantile classification method was used. Knowing the specific value of emissions per unit area is not particularly useful, thus being able to distinguish between the classes is not important. This allows the use of a larger number of classes for a more continuous visual.
4.2.3. Spatial Interpolations

Several Interpolation methods were used to explore the data. Inverse distance weighted, spline with tension, and ordinary kriging were tested and the results are provided below in Figure 9, Figure 10, Figure 11 respectively.

Figure 9 provides the output of the IDW interpolation method. In creating this interpolation raster, a P value (for use in Equation 7) of 2 was used. Although the RMSPE statistic returned a value of 1, when comparing the outputs, the default value of 2 provided a more detailed result. The spline interpolation method requires no input besides the raw data. Figure 10 provides a visual of the resulting interpolated raster surface from the spline method. The final interpolation presented is created using the ordinary kriging method. This one was chosen as the trends did not seem to vary, but the regression coefficient was unknown. After all the interpolations were obtained, the areas containing values in the top quartile or extracted and overlaid. Where the top quartile of each of the three interpolations intersect is presented in Figure 12.
Figure 9: Inverse Distance Weighted Interpolation of Yearly Emissions, Quantile Classification

Figure 10: Spline with Tension Interpolation of Yearly Emissions, Quantile Classification

Figure 11: Ordinary Kriging Interpolation of Yearly Emissions, Quantile Classification
4.2.4. Hot Spot Analysis
Incremental spatial autocorrelation was used to find the distance band, or neighborhood, that points are compared within. The distance that resulted in the highest intensity of clustering was at a radius of 1.6 km. Notice that clusters of low values occurred only within the inner city while clusters of high values were present in multiple communities outside the (inner) city.
4.2.5. Buffer Averaging

The choice in buffer size was a balance between perceptibility and resolution of the included data. The larger the radius of each buffer, the more visible it is. This comes with the tradeoff of more buffers overlapping, resulting in a visual that dilutes the values at both ends of the spectrum. After testing a few different sizes, a radius of 1250 meters was chosen for the overview of Uppsala municipality shown in Figure 14. This size seemed to provide a good balance of perceptibility and resolution. When zooming in or viewing an area in more detail, a smaller radius would be desirable. In kind, when zooming out or viewing a larger area, an increased radius would be desirable.

Although the mathematics behind this theory are simple, practical difficulties exist. Even with modern computers, physical memory and processing capabilities are limited; thus, the size of the data set used as an input for this method is limited. In order to complete this computation for the dataset used in this study, minimizing the number of raster cells averaged at one time was required. This was accomplished breaking the dataset into sections and calculating the average. These were then combined through a final calculation for the average of the entire area.

Figure 14: Buffer Average. Displays emissions from personal automobiles at the vehicle’s registered address. Values are given in CO\textsubscript{2}/year as calculated by yearly travel multiplied by car emission factor as stated in Transportsyrrelsens registry. Emissions are represented as equal sized circles and where emissions overlap, the mean is created.
5. **Discussion**

Uppsala has high ambition to achieve the goals of the Paris Agreement, higher than many others: become fossil-free by 2030 (WWF 2019). A large part of this goal comprises decarbonizing transportation within the area. To achieve transformation of an entire system within such a limited time, methods must be strategic and calculated. As there is no ‘one size fits all’ solution, a suite of policies that promotes co-modality is needed to address the all the different mobility patterns of residents. Clear communication and collaboration between data scientists and local politicians can result in evidence-based actions that efficiently transitions the transportation infrastructure. Findings for Uppsala municipality are presented in this section, but the same process can be followed for other municipalities.

5.1. **Putting It All Together: A Suite of Insights and Policies**

Recent actions taken by the Swedish government such as carbon taxes have had lackluster results (Klier & Linn 2015). This study aims to understand the reliance on ICEs and its context on a local level using GIS. This section discusses what insights the various visualization methods provide, various cohorts that are revealed, and explores policies that could strategically target each aforementioned cohort.

*Limitations of the Data*

The analyzed dataset exhibits approximately 0.177 vehicles per capita. Compared to the national and regional statistics of 0.477 and .437 vehicles per capita (Trafikanalys 2018b), indicating that between half and two thirds of the vehicles on the road are omitted from the analysis of this study. Although the sample size used is substantial and the missing datapoints were excluded for good reason, results may vary if valid information on the missing datapoints could be included. Future studies may have higher quality data available, resulting in a more comprehensive dataset.

It is worth mentioning the limitations regarding how this study accounts for emissions. Rather than studying emissions from a territorial perspective, a consumption perspective is applied. This allows to understand whose mobility patterns result in emissions and to what level. An understanding is gained from this into what neighborhoods can be targeted. Other studies conducted using a territorial perspective of where emissions are produced can offer insights into where commuters mobility demands currently exist.

*Distribution of emissions*

The QQ-plot in Figure 5 indicates that a small portion of vehicles is responsible for an unproportionally large amount of emissions. In an attempt to understand why such high distances are being driven by the high emitters, the locations of the top 1% and top 10% were studied in isolation. One might conjecture that high distances would be correlated with living in rural areas, however Figure 15 shows that the urban/rural mix of the top emitters is quite evenly split. Of the top 1%, approximately 40% live within Uppsala municipality. The top 10% is even more equally distributed, with essentially half residing within Uppsala municipality.
It has become increasingly common for nation states to institute taxes and subsidies on vehicle purchase and ownership based on its carbon emissions. These financial policies demonstrate some short-term effects in consumer patterns; however, do not stimulate changes over the long term (Klier & Linn 2015). Cities have begun to use taxes as well to modernize the vehicles used within the city limits. For example, London introduced a new tax this year charging ICE vehicles built before 2006 a £12.50 fee to enter the city to reduce smog (O’Sullivan 2019). Such taxation policies target the fuel efficiency of the vehicle which, as shown in Table 3, is less of a contributor than the distance driven. Rather than simply tweaking the status quo of automobile usage by encouraging the use of more efficient ICEs, financial incentives could be used to promote a modal shift by instituting tax policies on the distance driven. An exponential rate could be used to target the top emitters shown in Figure 15.

Current trends of taxes focusing on fuel efficiency may be due to the complexity of constructing policy based on distances driven. As previously discussed, the geographical distribution of the highest emitters is quite uniform. Those in rural areas may not have the availability of alternative modes as those in more urban areas have. Recent events such as the rise of the ‘Gilets Jaunes’ protests have taught us that “a fossil free economy will have to be a fairer economy” (Williamson 2018). Finesse would be required in the creation of taxes based on distance driven in order to avoid such issues. This could include relaxing the tax rate for those with less access to public transport and increasing it for those in urban areas where many alternatives exist.

**Choropleth**

The commonly used choropleth method was used in the first effort in understanding the data. The resulting visual presented in Figure 7 suggests that the closer a residence is to the urban center the lower the automobile emissions are. This of course makes sense and confirms a common understanding but...
introduces concerns of selection bias. Commuters that have longer commutes and poorer access to public transportation naturally tend to rely more heavily on automobiles. Urban and rural settings are undeniably two different circumstances and should be deliberated separately in fulfilling their mobility demands.

Choropleth maps are often criticized due to a frequent disregard of the geographical distribution of data. This leads to unproportionally representing the data. One strategy to avoid this mistake is to base the map on percentages or ratios (Kemp 2007). As Figure 7 does not quite follow this suggestion, a choropleth map using a ratio of yearly emissions per capita (the sum of all data points within a postal code divided by the population of said postal code) was created and is presented below in Figure 16. Analyzing the data in this way not only analyzes where emission levels are high, but also considers how the emissions are distributed between the population.

Figure 16 reveals a grouping of postal codes (specifically 75644, 75318, 75323, and 75450) have a much higher emissions per capita than other urban areas. These postal codes make up the Boland area, which contains many large shopping centers rather than being a residential area. The use of postal codes may isolate or differentiate areas of similar land usage; however, other visualizations validate the finding that the few residents located in Boländerna tend toward higher automobile emissions.

A small body of research shows that temporary interventions in habitual behavior are enough to cause a mode shift. Incentives such as promotional offers for public transportation or disincentives such as temporary road closures could produce large enough interventions to allow for the creation of other habits (Wynes, Nicholas, Zhao & Donner 2018). Since public transportation and cycle paths are readily available in the areas indicated by the choropleth map, instituting informational campaigns could increase usership. Inspiration could be taken from existing campaigns such as the ‘winter cyclist challenge’, where 26 students are provided with equipment that makes winter cycling safer in exchange for their service as advocates. Their advocating includes setting an example through cycling throughout the winter as well as participating at various events where they spread information on safe bicycling and hand out lights to fellow students (Meyer-Rodrigues 2019). Instead of targeting a person with a student status, residents of a certain geographical location could be targeted.
Kernel Density
Kernel Density was the second method used to analyze the data. This allowed to look at the data in more exact terms than postal code areas. A kernel density map, provided in Figure 8, shows the total yearly emissions of each square raster cell.

Initial impressions were that this method largely reflects population densities. This hunch was confirmed by comparing the results from the kernel density with population data (Figure 17). To remove population density from the visual, Figure 18 provides the result of dividing the kernel density by a raster of population obtained from SCB (2019).

The resulting image provides two insights, firstly, the same area of interest indicated in Figure 16 appears as a high emissions area in Figure 23. Secondly, suburbs including Läby, Björklinge, Skuttlunge, Vattholma, and Storvreta, are indicated as areas with higher emissions per capita. The higher emissions areas are not at the center of these suburbs, most likely due to public transportation being located here, but generally a kilometer or two away from the suburbs’ center. These areas could be reached by increasing accessibility to the bus stops with strategies such as constructing or improved cycle paths and providing bicycle parking at the bus stops. Alternatively, modes such as carpooling could be promoted.
In an attempt to consider missing data points, several interpolation methods were used. The basic idea of these methods is to use a mathematical formula to fit a surface to the existing data points. This estimates the values in between the existing data points. Several interpolation methods were utilized, and the results compared. If a trend did not appear in each of the interpolations, then the instance was considered an anomaly, disregarded, and attributed to an incongruity of that specific mathematical formula.

The interpolations confirmed the general trend revealed by the choropleth map that residencies further away from the urban center often have higher emissions. The interpolations are useful as they provided a higher resolution than the choropleth map of this trend and decouple the visual from population density. Deciding the extent that population density can be complex. While results from kernel density skew the data towards areas of high population density, the interpolations have difficulty producing realistic visuals of rural areas. A portion of the indicated high emissions areas are uninhabited. Interpolation methods may provide the most benefit in analyzing trends in densely populated areas.
**Hot Spot Analysis**
Spatial statistical analysis provides a different lens to view the data through. As explained in the methods section in greater detail, hot spot analysis is based on statistics and indicates nonconformity within localized clusters of data points. This method provides an easily understandable visual of the statistical relevance of individual data points.

Applied in this study, it clearly demonstrated the trend previously seen in the choropleth and interpolations, of residents close to the urban center had lower emissions from automobile usage while those residing further away produce higher emissions. As previously discussed, this phenomenon indicates selection bias. Population selection can become problematic when analyzing both rural and urban areas.

More interestingly, this map also confirms the findings from the choropleth map presented in Figure 16 by revealing hot spots in the same area.

**Buffer Averaging**
A custom script was written to create a visual that provides both the ease of readability of a gradient as seen in the heat map while still allowing details such as those shown in the hot spot analysis to be understood.

Many benefits are provided by this method. While it also interpolates space between points to a limited extent, it creates a more accurate output than the other interpolation methods presented and shows null values for uninhabited areas. It solves the issue of emphasizing high concentrations of datapoints as is seen in kernel density. However, the other methods provide value as outliers have high influence in data scarce areas while are not visible in dense areas. This could be solved through refining the script take into account point density by utilizing percentages or ratios as was done previously with the choropleth map. This method will be used for further examination in the following section.

**5.2. A Deeper Dive into Evidence Based Policy: Developing Public Transportation as an Example**
To provide an example of how emissions data could be used to drive decisions, accessibility to the mass transit system was evaluated. Public transportation is already widely available within Uppsala due to the extensive network of established bus routes. To future-proof the mass transit system, the municipality is considering installing a tram system in Uppsala’s city center (Uppsala kommun 2019).

Uppsala’s tram project targets the areas with a high population density which have a high positive correlation to the ‘hot spots’ shown in the kernel density visualization of Figure 8. Increasing capacity for the growing urban population is important, but other initiatives targeting larger emitters or rural areas should also be considered. While methodologies exist for collecting information on where and when commuters use existing infrastructure, such as mining smart cards to evaluate bus lines (Yu & He 2017), it is more difficult to study where the demand is not being met. This section provides a strategy for doing just that, with the results presented in Figure 19 through Figure 22.

To analyze the public transportation network, the visuals from previous sections were overlain with a visualization of the accessibility of public transportation. The existing bus lines as well as the current proposal for the tram line. Accessibility for the bus lines is shown by placing a buffer around each bus stop. As stops have not been proposed for the tram line yet, a buffer was placed around the entire line. A distance of 400 meters is frequently used as a reasonable distance for a commuter to walk in an urban context to gain access to public transportation (Murray & Wu 2003); thus, this distance was used as the buffer radius.
While many of the larger emitters are within a reasonable distance to the bus stops, it can be seen that a few areas would benefit from an expansion of the bus network. Other areas contain such a small populace that the addition of a bus stop may prove impractical. In these cases, strategies which increase the accessibility of surrounding bus stops could increase ridership. Providing areas to park bicycle at bus stations and ensuring cycle paths exist and are maintained could contribute to increasing the distance commuters are willing to travel to bus stops. Informational campaigns could convince rural residents to carpool together into areas where other modes are more accessible.
Figure 21: Comparison of Kernel Density automobile emission visualization with the proposed tram route.

Figure 22: Comparison of Buffer Average automobile emission visualization with the proposed tram route. A reduced radius of 500m is used for the buffers in the displayed buffer analysis for increased resolution in this ‘zoomed-in’ map.
Granted many factors must be taken into consideration when planning the tram route, emissions data presented in Figure 21 and Figure 22 show that much of northern Uppsala’s urban area is not addressed by the current plan. Similarly, there appears to be heavy automobile reliance in Sunnersta that access to a tram line could reduce by extending the Ultuna route or adding a third loop. Eriksberg is another densely populated area that is not addressed by the tram line plans; however, the average emissions are already relatively low which might signal other modes such as cycling are more common.

5.3. Possible Areas for Improvement and Recommendations for Further Studies

Possibly the largest weakness of this study is the incomplete data set used. Although the results produced above are not invalid, a complete dataset may reveal additional areas of interest. Without the restriction of time, missing values of CO₂ emissions could be pulled from a database based on the make, model, and year of the vehicle. This process could be eased by automating the process with another python script.

A detail to keep in mind when compiling emissions data is the standards that are used in measuring the said value should be consistent throughout the dataset. With the New European Driving Cycle (NEDC) standard being phased out and replaced by the Worldwide Harmonised Light vehicle Test Procedure (WLTP) standard, an understanding of which standard is used and the implications for using different standards is important. Transportstyrelsen is clearly aware of both NEDC and WLTP emission testing standards as the data obtained from them includes fields for both; however, all values for WLTP remain blank. Transitioning suddenly to the new standard could provide results more representative of real world driving but using a mix of the two would be ill advised. The discrepancy between laboratory test results and actual performance has increased in Europe has increased over the last two decades (Willerström 2019). Test results from other continents could be looked to for obtaining more accurate data (Hooftman, Messagie, Van Mierlo & Coosemans 2018).

Furthermore, this study makes assumptions regarding driving style. Emission values for urban, rural, and urban/rural mixed driving styles based on the NEDC were available. As not enough information is available to determine which driving style is most applicable for each data point, this study applies the same driving style to all data points. Thus, the NEDC value for mixed urban/rural is used in Equation 1 for each data point. The accuracy of the results could be improved if a rational technique was available for assigning emission values based on the appropriate driving style to each data point. For example, a residence in a rural setting would be required to spend a much larger portion of the commute in a rural setting, thus the NEDC value for rural driving may be more representative.

The results presented in this report only consider one point in time, resulting in a ‘snap-shots’ of the situation. Once emission mitigation policies have been implemented, revisiting the data is vital in analyzing if the mitigation strategies are achieving the desired results. Analyzing trends over time could be done by producing the visuals outlined in this study at regular time intervals. This could be done through creating animations from the results or using them to create new still images where the difference in value is presented (Kemp 2007). For example, areas with a decrease in emission would correlate with negative values with the reverse for increasing emissions and the magnitude would represent the rate of change. These values would then be color coded for the visual in a similar manor as done within this report. For example, red may be chosen for the areas of increasing emissions and blue or green chosen for areas of decreasing emissions. Several of the map styles presented here could be adapted to visualize this.

The results of this report could be further analyzed under two different contexts: urban and rural. Analyzing these two different settings of differing mobility needs separately would prevent any issues with selection bias. Some visualizations had clearer issues with the population selection problem than others, but Section 5.2 attempted to addresses this by separately addressing co-modality in urban and rural
areas. Under the urban context, accessibility to mass transit is discussed in terms of capacity and how a tram line can maintain co-modality in a growing population. In contrast, accessibility to public transportation is discussed in the rural context in terms of a lack of infrastructure, causing the automobile to be more convenient and thus more relied upon.

Regarding the analysis of the public transportation network, a Euclidian distance was used in assessing accessibility, which produces inaccuracies due to not accounting for variables such as difficult terrain or nonlinear walking/cycle paths (Zhao, Chow, Li, Ubaka & Gan 2003). Future analysis could be done using more accurate methods for analyzing accessibility. Just as bus stops were overlaid with automobile emission data, bicycle paths could be investigated in a similar method in regard to reducing automobile emissions. Further studies may find interest in studying more in-depth Uppsala’s tram line proposal by evaluating the factors considered and how they align to Uppsala municipality’s agenda 2030 goals and the SDGs.

Other sources could provide valuable data. Mobile phones have become widespread over the last decade. Access to call detail records – data generated every time a call, text, or the internet is accessed via a mobile phone that logs the time, type of communication, and cellular tower used – make analyzing mobility on a micro level possible (Williams, Thomas, Dunbar, Eagle & Dobra 2015). This would allow analysis on multiple modes, shorter temporal periods, and a better understanding of the ambient populace.

The homogenized use of mobile phones has led to a rise in mobility as a service (MaaS). Ridesharing companies providing access to bicycles, electric scooters, and ride hailing through mobile apps have become increasingly popular. A private-public partnership would benefit both city planners and private companies in promoting co-modality. With an open data policy, private actors would be able to integrate public transportation, making their services even more convenient. In exchange, city planners would have access to highly detailed information of mobility patterns and behaviors generated by the on demand MaaS platforms (Barreto, Amaral & Baltazar 2018). With such a wealth of data, regression analysis could be performed to better understand what variables most contribute to the commuter’s choice in co-modal transport over reliance on automobiles.

In addition to using other data sources, further research may use other methods of analysis to create a deeper understanding of the data. Many forms of regression analysis are available to examine and explain factors driving spatial patterns. For example, choice modeling could be used to model and predict the impacts of economic factors in public transportation usage.
6. Conclusion

The transport sector requires radical transformation in order to create a sustainable environment we can all live in. For this transformation to be sustainable, strategies need to consider social, economic, and environmental factors. Many national governments have made promises and set goals for reducing or even eliminating emissions from the transport sector. Many of these efforts have resulted in incommensurate results. This study works to understand and communicate geographical aspects of emissions on a local level to establish a method for municipalities to create evidence-based mitigation policy. The city of Uppsala was used as a case study due to the city’s ambitious environmental goals.

The case study of Uppsala was used to demonstrate how visuals created with GIS can be used to both create an understanding of and communicate automobile emissions data. This provides a foundation for creating evidence-based emissions mitigation policy. Vehicle registration data was obtained from Transportstyrelsen, geocoded, imported into the GIS program, statistically analyzed, and finally used to create visualizations.

Four main insights were found in the case study of automobile emissions within Uppsala municipality. Firstly, the geographical distribution of the top percentile of emitters is relatively homogeneous. Although further research into socioeconomic factors that these emitters have in common could prove interesting, simply identifying them and their location is enough to begin creating policies that target them. Second, urban areas have larger total emission than rural areas due to large population densities. Rural areas have higher per capita emissions than urban areas due to the availability of fewer alternatives to the automobile and needs to travel longer distances. Lastly, Boländerna was a prime example of how areas within a city can rely heavily on automobiles even though other modes are readily available. Different strategies for rural and urban contexts were explored for how access to public transportation can be improved in order to promote co-modal mobility; however, future studies may provide more clarity in mobility patterns by studying rural and urban emissions separately.

Perhaps the most interesting part of the report is how visualizations of emissions can be used to develop climate mitigation policy. Emission visualizations were superimposed with geographical data of both existing bus stations and proposed tram route. Buffers were placed around these to estimate the accessibility of Uppsala’s mass transit. Combining information on the accessibility of buses and knowledge of which addresses were associated with high automobile emissions enabled identifying locations where the bus network could be improved. Strategies for increasing the effectivity of the public transportation network were then explored. Building out infrastructure is needed in some cases where public transportation is not available, or access to it is poor. Where mass transit is already available, policies aimed at influencing behavior to induce a mode shift are more pertinent. Strategies could include informational campaigns, temporary financial incentives such as a free trial bus pass, or disrupting habitual automobile usage through placing temporary obstacles such as road closures.

The future of GIS contains much potential in developing evidence-based climate mitigation policy of urban and rural planning. This study lays out a methodology that can be built upon. While automobile registration data was used here, an increasing number of sources are becoming available.
7. Acknowledgements
I would like to thank Martin Wetterstedt for acting as supervisor for my thesis and offering his guidance. His feedback always provided great clarity and inspiration. I’d also like to thank John Östh for reviewing my thesis and providing his expertise on GIS. I thank Alan Sarits for providing an exceedingly thorough opposition and his insights from a social science perspective.

I have much gratitude for my incredible classmates and people I have met throughout my time in this program. Overcoming obstacles together, joining in laughter, and sharing wisdom has been invaluable. I have much gratitude for friends, new and old, as well as family for providing unwavering support and belief in me.

Much of the thesis was written in Uppsala’s wonderful cafes. I am beholden to the baristas who fueling me with coffee; however, I blame any misspellings or grammatical errors on the caffeine induced jitters.

When everyone else failed me, the Stack Exchange and Stack Overflow communities were there. I am forever indebted to the wisdom of the forgotten usernames that helped me in coding my python script when the error messages were unhelpfully ambiguous.

Lastly, I’d like to thank the Swedish government for collecting data on its populace and for each individual for following the law in registering their personal vehicles, making this study possible. In a related vein, I’d like to acknowledge the fossil fuel industry for their hard work in making the topic of this thesis relevant.
8. **References**


Appendix A. Data Security
To ensure security of the data entrusted to the researcher by Transportstyrelsen, measures will be taken to prevent the data from spreading. During the course of this project, the data

a) will be stored on a secure local drive and not saved to the cloud,
b) will not be loaded onto public computers
c) will not be loaded onto USB drives that can easily be lost or misplaced
d) will be shared through encrypted and or password protected methods
e) will be limited to only the necessary data (i.e. only addresses) when using ESRI’s online services
Appendix B. Distribution of data in visualizations due to classification methods

Figure 23: Frequency distribution of data in the Choropleth map of Figure 7 with the quantile classification method employed

Figure 24: Frequency distribution of data in the Choropleth map of Figure 16 with the quantile classification method employed
Figure 25: Frequency distribution of data in the IDW map as presented in Figure 9 using the quantile classification method. Frequency distribution for the Spline with Tension as displayed in Figure 10 was omitted due to no perceptible difference from the distribution of the IDW method.
Figure 26: Frequency distribution of data in the Ordinary Kriging as shown in Figure 11. A manual classification method was used to match the class ranges of the IDW and Spline Tension methods to ease comparison between the methods.

Figure 27: Frequency distribution of Hot Spot Analysis as shown in Figure 13. ‘-3’, ‘-2’, and ‘-1’ correspond to 99%, 95%, and 90% confidence of cold spots, respectively. The same goes for the hot spots, but these values are positive.
Figure 28: Frequency distribution of the buffer analysis as shown in Figure 14. Standard deviation is used here as the classification method with an interval size of ½ standard deviation.

Figure 29: Frequency distribution of the buffer analysis as shown in Figure 20Figure 14. Quantile distribution is used here as the classification method.
Figure 30: Frequency distribution of the buffer analysis as shown in Figure 22. Quantile distribution is used here as the classification method.
Appendix C. Buffer Averaging Script Logic

1. Start
2. Check ArcGIS extensions
   - Not Available: Return Error → Stop
   - Available: Import emissions data → Find unique values → Extract emission value
3. Buffer Radius
4. Create buffer for value → Save buffer to temporary raster
5. Print Progress
   - No: # buffers = # values? Yes → Print ‘Buffer conversion done’ → Calculate cell statistics (average) of overlapping buffers → Print Progress
   - No: # calculated chunks = # total chunks? Yes → Print ‘Chunks completed’ → Save chunked average to temporary raster → Stop
   - No → Save temporary files → Print ‘Temp files deleted’

   - All cells of ‘chunk’ calculated → Print Progress
     - No: # of chunks → 'Chunk' Size
       - No → Raster Cell Size
         - No: ‘Chunk’ Size → Calculate cell statistics (average) of overlapping buffers → All cells of ‘chunk’ calculated
         - Yes → Print ‘Chunk done’ → Stop
     - Yes: ‘Chunk’ Size → Calculate cell statistics (average) of overlapping buffers → All cells of ‘chunk’ calculated

Appendix D. Buffer Averaging Python Script

1. # This script processes a shapefile including multiple polygons (such as buffers) with
2. # unique values. These polygon features often overlap. The script takes each feature
3. # in the shape file and exports each feature individually. This provides a library of
4. # temporary shapefiles including only one polygon and its respective attributes.
5. # The temporary shapefiles are then converted to a raster. Finally, all the individual
6. # rasters are averaged where they overlap. To make the process of averaging all the
7. # individual rasters less memory intensive, the process is broken down into smaller
8. # “chunks”. The “chunks” are then averaged for the final output. This script requires a
9. # specific folder structure and automatically clears temporary data
10.
11.
12.
13. #function for finding unique values of a field in FC
14. def unique_values_in_table(table, field):
15.     data = arcpy.da.TableToNumPyArray(table, (field))
16.     return numpy.unique(data)
17.
18. #function for breaking rasterList into more manageable (less memory intensive) pieces
19. def chunks(l,n):
20.     #yield successive n sized chunks
21.     for i in xrange(0, len(l), n):
22.         yield l[i:i+n]
23.
24. import arcpy, numpy, os, gc
25. from arcpy.sa import *
26. from arcpy import env
27.
28. #check extensions
29. try:
30.     if arcpy.CheckExtension("Spatial") == "Available":
31.         arcpy.CheckOutExtension("Spatial")
32.     else:
33.         # Raise a custom exception
34.         raise LicenseError
35.     except LicenseError:
36.         print "spatial Analyst license is unavailable"
37.     except:
38.         print arcpy.GetMessages(2)
39. finally:
40.     #Check in the 3D Analyst extension
41.     arcpy.CheckInExtension("Spatial")
42.     #parameters and environment
43.     temp_folder = r"C:\Users\Joran\Documents\MSDThesis\TEMP\EmissionMaps\tempRasters"
44.     output_folder = r"C:\Users\Joran\Documents\MSDThesis\TEMP\EmissionMaps\outputRasters"
45.     env.workspace = temp_folder
46.     unique_field = "FID"
47.     #specify field to be analyzed
48.     field_of_Interest = "EmissionsC"
49.     #specify cell size (smaller = greater resolution)
50.     cellSize = 100
51.     #specify buffer file
52.     print "importing data"
53.     fc = r'"C:\Users\Joran\Documents\MSDThesis\TEMP\BufferExports\r2500m.shp'"
54.     stat_output_name = fc[54:-4] + ".tif"
55.     #All features or cells will be processed.
56.     arcpy.env.extent = "MAXOF"
make layer for selecting

arcpy.MakeFeatureLayer_management (fc, "lyr")
totFeatures = arcpy.GetCount_management(fc)

#go through each row of the buffer dataset and create a separate raster file for each row
SC = arcpy.SearchCursor(fc)
for row in SC:
    object = row.getValue(unique_field)
    value = row.getValue(field_of_Interest)
    expression = 'FID' = ' + str(object) + '
    #select the feature
    arcpy.SelectLayerByAttribute_management("Lyr", "NEW_SELECTION", expression)
    #create feature class name, with the row number at the end of the name
    outFeatureClass = "Buffer_" + str(object) + ".shp"
    #export the feature
    arcpy.FeatureClassToFeatureClass_conversion("lyr", temp_folder, outFeatureClass)
    #print status of exporting loop
    print "exported FID " + str(object+1) + " / " + str(totFeatures)
#remove the temporary buffer shape files
arcpy.Delete_management(outFeatureClass)
print "conversion done"
gc.collect()

#check out the ArcGIS Spatial Analyst extension license
arcpy.CheckOutExtension("Spatial")

to run large datasets on machines with limited memory, the dataset is broken into smaller "chunks" and averaged
print "chunking"
chunkSize = 500
totChunks = -(num_of_rasters//chunkSize)
print "number of chunks: " + str(totChunks)
tempCount = 1
for rasterList_Chunk in chunks(rasterList, chunkSize):
    print "processing TempCellStats " + str(tempCount) + " / " + str(totChunks)
    TempCellStats = CellStatistics(rasterList_Chunk, "MEAN", "DATA")
    temp_fileName = "chunkedCellStats" + str(tempCount) + ".tif"
    temp_output_name = os.path.join(temp_folder,temp_fileName)
    TempCellStats.save(temp_output_name)
    tempCount = tempCount + 1

#create list of temporary raster chunks
print "Compiling Temporary Chunks"
chunkRasterList = arcpy.ListRasters("chunked*","ALL")
chunkRasterList.sort(key=lambda f: int(f[16:-4]))
gc.collect()
#take an average of all the rasters

```
print "Calculating Average"
outCellStats = CellStatistics(chunkRasterList, "MEAN", "DATA")
print "saving output" + stat_output_name
stat_output_name = os.path.join(output_folder, stat_output_name)
outCellStats.save(stat_output_name)
print "save complete"
```

#delete chunked rasters
```
print "deleting chunked rasters"
print chunkRasterList
for row in chunkRasterList:
    count = int(row[16:-4])
    print "deleting chunk " + str(count) + "/" + str(totChunks)
    arcpy.Delete_management(row)
    arcpy.Delete_management("lyr")
```

#delete the remaining temporary rasters
```
print "deleting rasters"
for row in rasterList:
    count = int(row[7:-4]) + 1
    print "deleting buffer " + str(count) + "/" + str(totFeatures)
    arcpy.Delete_management(row)
    arcpy.Delete_management("lyr")
arcpy.CheckInExtension("Spatial")
```