What role does urbanization, using China as an example, play in shaping average temperature trends in a warming world?

YANG YU, University of Toronto, Canada

In this study, we analyze the relationship between urbanization and average temperature trends in China. Using temperature data and various urbanization indicators, we investigate how urbanization has contributed to the warming of China over time. Our analysis shows that urbanization has played a significant role in shaping the average temperature in China, and the effect varies between the north and south regions. The south region has experienced more extreme low-temperature days and a decrease in average temperature from 1950 to 2012, while the north region has experienced more high-temperature days and an increase in average temperature during the same period. These trends could be attributed to various urbanization effects, such as industrialization in the north region during 1950-1978. To better understand the relationship between urbanization and temperature, we conduct regression and machine learning analyses and find that population has a positive impact on increasing the national temperature. Our findings provide valuable insights into the complex relationship between urbanization and climate change in China.

Author's address: Yang Yu, yangy.yu@mail.utoronto.ca, University of Toronto, 50 St George St, Toronto, Ontario, Canada, M4Y1G2.

XXXX-XXXX/2023/4-ART $15.00

https://doi.org/10.1145/nnnnnnn.nnnnnnn

Additional Key Words and Phrases: temperature change, global warming, urbanization, China

ACM Reference Format:

Yang Yu. 2023. What role does urbanization, using China as an example, play in shaping average temperature trends in a warming world?. 1, 1 (April 2023), 30 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

With the development of modern cities and urbanization, the global average temperature has been increasing over the past century, leading to concerns about the impacts of climate change. While urbanization significantly improves life standards and society efficiency, it is commonly believed to be a significant contributor to the global warming process.

China, one of the fastest-growing developing countries in the world, is undergoing rapid urbanization in the recent few decades. However, the rapid urbanization also comes with high pollution level in China and serious environment problems. Because of this, researchers have studied how the urbanization process affect the environment and how does it contribute to the high pollution in China. It is commonly believed that although the rapid urbanization is an indicator of China's economic prosperity. Li et al (2021) suggests that living in the cities and the urbanization effect has a positive effect on subjective well-being and improve their stratification. It is a unknown question of how the urbanization will impact the environment. Global warming effect has been the center of the discussion. There are a lot of evidence from the world showing that the world is becoming warmer, due to the greenhouse effect, which is the result from urbanization and industrialization. Although we have abundant evidence from the developed country such as the United States or other European countries showing that the urbanization process do increase the speed of temperature change, this question is still unclear for developing countries like China, which started their urbanization process almost a century later than developed countries. This research aims to examine the role of urbanization in shaping average temperature trends in China based on a time series analysis starting from 1950 to 2012.

Specifically, 1950 to 1978 is the period of Chairman Mao Zedong's era, named after the founder of the Communist Party of China. Economic development was unstable and slow due to some policies that were originally aimed to achieve the goal of industrialization. The process of urbanization and modernization was slow. 1979 to 2012 is the period China decided to begin the journey of reforming and opening up, also known as the period of the Deng Xiaoping era, named after the second leader that achieved economic recovery and development. He proposed a series of policies that aimed to accelerate the process of industrialization, including one of the most important policies in Chinese history: the Chinese reform.

The aim of this analysis is to investigate the relationship between average temperature and several independent variables, including GDP, urban and rural population, sectoral employment, electricity generation, and other measures of urbanization in China from 1950 to 2012. We first examine the trend of annual temperature change over the entire period and find that the average temperature is increasing, with the period from 1978 to 2012 showing the most significant increase, aligning with the highest economic development speed. We then analyze the frequency of cities that experience temperatures over a certain value using percentage of annual average temperature as the independent variable. We also investigate the relationship between annual average temperature and urban population, observing a positive correlation between urban population and average temperature, as well as a positive relationship between the percentage of the total population residing in urban areas and average temperature in China over the entire period. In addition, we utilize heat maps to visualize the change in average temperature across China.

Several studies have investigated the role of urbanization in shaping average temperature trends in China. Li et al. (2013) highlighted the importance of land cover change and human activity in contributing to warming temperature in China. As a result of urbanization, nearly half of the chosen stations have been converted to urban stations due to land cover change. In urbanized station, they observed the most warming effect. Zhang et al. (2010) states that the conversion of rural land to urban land cover results in significant changes to near-surface temperature, humidity, wind speed and precipitation. Similarly, Sun et al states that the contribution of urbanization to warming in China was found to be significant, accounting for approximately one-third of the total warming (2016). In addition, urbanization does not only affect warming temperature due to economic development and industrialization, but also on its population scale and geographic environment (Fang et al., 2013). Similar to our approach, the researchers divide the 49-year study period into two periods: a period of less economic development, from 1961 to 1978, and a period of greater economic development, from 1979 to 2009. They found out that the population scale of a city represents a significant factor: a city with a larger population has a stronger warming influence, regardless of whether the effect is negative or positive. This could be helpful for us to conduct future research on the relationship between population and temperature change in China.

The findings of our study reveal an overall increase in average temperature in China from 1978 to 2012, while temperatures remained stable from 1950 to 1978. Furthermore, our results demonstrate a positive relationship between urban population and average temperature, indicating that urbanization and economic development are positively associated with warming in China. Notably, the south of China has experienced a decrease in average temperature, while the north has seen an increase in average temperature, suggesting a shift in temperature patterns across the country. This finding is consistent with the findings of Ren et al. (2008), which stated that the regional average annual mean temperature series in north China is significantly impacted by urban warming. Our study also suggests that global warming may contribute to more extreme temperature years in China. In addition, we observed a positive relationship between the percentage of people working in the service sector and the change in average temperature, reflecting China's transition from an agriculture-based economy to a service-based economy.

In addition to the findings from the graphs and evidence, our study also utilizes both multiple regression and regression tree methods to gain a deeper understanding of the relationship between urbanization and temperature trends in China. Our regression analysis reveals that population growth and urbanization are likely contributing factors to the rise in average temperature in China. Surprisingly, the results indicate that electricity generation is negatively correlated with temperature, which was unexpected. However, our regression tree analysis provides additional insights by revealing that provinces with higher electricity generation growth rates generally experience higher temperatures, consistent with our initial hypothesis.

Furthermore, we explore the difference in the effect of population on temperature between the north and south regions of China. Our findings suggest that provinces with higher population growth rates are generally more prone to experiencing elevated temperatures, regardless of geographic location. Provinces in the northern region of China with a population growth rate exceeding 8.36 percent are expected to experience a temperature rise of approximately 5.5 degrees Celsius, while provinces in the southern region with a population growth rate surpassing 7.74 percent will experience a temperature increase of around 2.29 degrees Celsius. By investigating the impact of urbanization on temperature trends in China, our study provides valuable insights into the complex interaction between human activity and the environment.

2 DATA

The data is collected from Kaggle to obtain monthly observations of average land temperature for major cities and countries globally from 1973 to 2013, including China. We choose to focus on China and use data starting from 1950, as this marks the founding year of the People's Republic of China, and the country was primarily an agriculture-based economy with low urbanization rates at that time. We believe that the high economic development and urbanization process since then make China a suitable case to explore the relationship between urbanization and temperature change.

From the World Bank Group, We also incorporate a new dataset that records the average temperature in China by provinces from 1901 to 2021. There are several reasons why we need to incorporate another temperature dataset. First, we found that the Kaggle dataset provided significant discrepancies with the map data due to outdated city names and the lack of inclusion of many cities in China based on the new Chinese city classification system. When we create map showing the average temperature in China, we find out that our shape file does not match the cities name in our Kaggle dataset and provide inconsistent map. Therefore, we decided to use the World Bank Group dataset as the first layer of the map to ensure the most accurate representation of the average temperature in China by province. We will use the Kaggle dataset as a second layer to display a more detailed average temperature map of different cities in China. Second, when we conduct our regression in the later part, most of our data are recorded in a province level or national level, not city level. Unlike developed countries where they have complete data, China does not have the complete data recorded in city level until 2000, which hinders our research if we continue to use the Kaggle temperature dataset.

To further investigate the urbanization trend in China, we plan to incorporate a new dataset from Kaggle that records China's population data. This dataset records the annual urban and rural population in China from 1950 to 2022, along with other relevant demographic information in a national level. This additional data will help us further investigate the urbanization trend in China, as population growth is a significant driver of urbanization. Unfortunately, our analysis is hindered by missing data for both urban and rural population in China from 1950 to 1959. Simply using a linear regression between years after 1959 and urban or rural population will not work, as it will overpredict the population for 1950 to 1959. An alternative approach would be to use a time series prediction model like the exponential smoothing model. However, applying this model requires a deep understanding and strong application. Therefore, we propose a simpler solution: to use the data of urban and rural population percentage of the total population for 1960 to fill in the missing values from 1950 to 1959. Since the variation in urban and rural population during this time is negligible, we can ignore the changes in population given the broader time range from 1950 to 2012. Furthermore, this period of time (1950-1978) is characterized by slow economic growth and urbanization in China, making it a suitable candidate for this approach. By filling in the missing values with the data from 1960, we can continue our analysis and explore the relationship between urbanization and temperature increase in China.

In addition, we decided to supplement the World Bank Group temperature dataset with another new dataset that records China's GDP by province from 1949 to 2020. GDP is a significant indicator of economic and social development, as well as urbanization. By doing so, we can visualize the economic development in China using a map, providing us with valuable insights into the country's progress over the years.

To enhance our analysis, we plan to use web scrapping to get the data records the total number of workers by sector. This dataset will help us determine trends in the number of people working in various sectors and their relationship with urbanization and economic development from 1952 to 2011. We can use line plots to show the trend of people working in different sectors and scatter plots to demonstrate the correlation between workers in specific industries and temperature change. Although not immediately apparent, the number of people working in different sectors is a strong indicator of economic development and urbanization, providing valuable insights to answer our research question.

China statistical yearbook provides the most comprehensive and accurate data from 1949 to the present. We decide to merge new datasets that records the electricity generation, population, number of vehicles, and highway mileage by province in China from 1950 to 2012. These datasets also provides valuable insight into the urbanization process across different provinces and how it correlates with economic development. We will further use these datasets to conduct regressions and data visualizations.

3 VISUALIZATION AND SUMMARY STATISTICS

The summary statistics table provides an overview of the coefficients for various variables in our dataset, which aims to investigate the role of urbanization in shaping average temperature trends in a warming world, using China as an example. The dataset includes 1461 observations, representing every province in China from 1950 to 2012. The observations exclude any missing data.

The table is divided into two sections. The first section presents summary statistics for the total number of vehicles (in 10,000s) and highway mileage (in km), which can serve as a proxy for urbanization levels. The dataset shows that the mean number of vehicles is 508,600 and the mean highway mileage is 44,631.24 km. The minimum number of vehicles is 46, indicating a lack of urbanization in some areas of China in the early years of the dataset. The maximum number of vehicles is 10,374,217, indicating a significant increase in urbanization over the years.

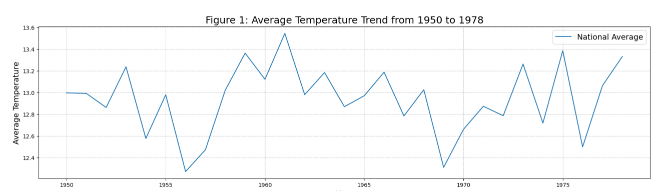
The second section of the table presents summary statistics for electricity generation (in 100m kWh), population (in 10,000s), average temperature, and GDP. The dataset shows that the mean electricity generation is 36,309 million kWh, the mean population is 35,352,700, the mean average temperature is 12.01 Celsius degrees, and the mean GDP is 2,870.58 billion yuan. The minimum electricity generation is 0.62 million kWh, the minimum population is 1,424,200, the minimum average temperature is -3.01 Celsius degrees, and the minimum GDP is 1.34 billion yuan. The maximum electricity generation is 400,113 million kWh, the maximum population is 1,143,000,000, the maximum average temperature is 25.23 Celsius degrees, and the maximum GDP is 57,067.92 billion yuan.

	Total number of vehicles(in 10,000s)	Highway mileage
count	1461	1461
mean	50.86	44,631.24
std	108.07	46,974.96
min	0.0046	1155
25%	1.7416	16,370
50%	11.02	30,298
75%	47.21	49,928
max	1037.4217	293,499

	Population(in 10,000s)	Average Temperature	GDP
count	1461	1461	1461
mean	3,535.27	12.01	2,870.58
std	2,305.04	5.99	6,286.17
min	142.42	-3.01	1.34
25%	1,765.9	7.79	64.07
50%	3,270.2	13.37	328.22
75%	4,857	16.25	2,587.72
max	11,430	25.23	57,067.92

3.1 Average Temperature from 1950 to 2012

We will first examine the average temperature in China using several graphs. Figure 1 uses line plots to demonstrate how the average temperature changed from 1950 to 2013 nationally. Specifically, we will divide the time interval into two periods: 1950 to 1978 and 1979 to 2012. The first period is the founding of the People's Republic of China, whereas the second period is China decided to begin the journey of reforming and opening up. The second period represents the urbanization and fast economic development of China. We will also see the overall picture from 1950 to 2012. From Figure 1, titled 'Average Temperature Trend from 1950 to 1978', we can see that the average temperature trend in China is relatively steady and is moving around at 13.0 degrees. This is the period of Chairman Mao Zedong's era, named after the founder of the Communist Party of China. Economic development was unstable and slow due to some policies that were originally aimed to achieve the goal of industrialization. The process of urbanization and modernization was slow. From the Figure 2, we can see that the average temperature trend in China is increasing. This is the period of the Deng Xiaoping era, named after the second leader that achieved economic recovery and development. He proposed a series of policies that aimed to accelerate the process of industrialization, including one of the most important policies in Chinese history: the Chinese reform. From the Figure 3, we can see that the average temperature trend in China is increasing. This graph provides us with a clear image of how the average temperature change in China from 1950 to 2012. The temperature is positively correlated with the development of China, including economic development and urbanization.



We further use bar charts to illustrate the trend of the number of cities with annual temperatures exceeding 23 degrees and below

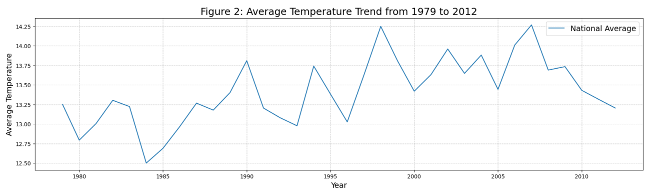


Figure 2: Average Temperature Trend from 1979 to 2012

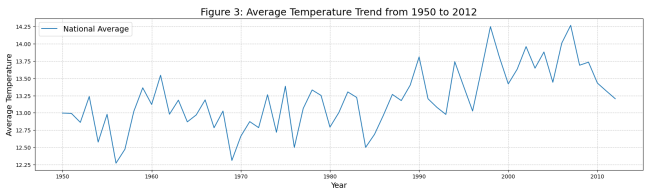


Figure 3: Average Temperature Trend from 1950 to 2012

0 degrees, over a 5-year interval from 1950 to 2010. Figure 4 indicates an upward trend, which suggests that the number of cities with an annual temperature of 23 degrees or higher has increased steadily over time. This trend is particularly prominent after 1980, coinciding with China's urbanization and economic growth. The graph provides further evidence that rapid urbanization has contributed to the accelerated increase in temperature. Figure 5 shows a decreasing trend, which implies that fewer cities have an annual temperature of 0 degrees or lower. This suggests that all cities in China are experiencing warming temperatures, rather than cooling. Once again, the graph highlights the impact of urbanization and economic growth on temperature changes in China, as cities become increasingly hotter.

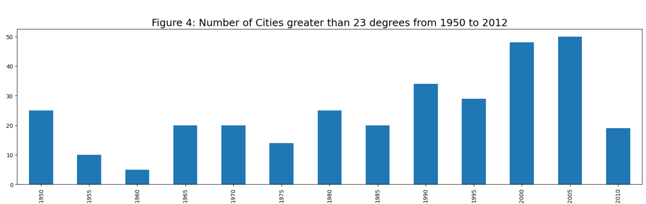


Figure 4: Number of Cities greater than 23 degrees from 1950 to 2012

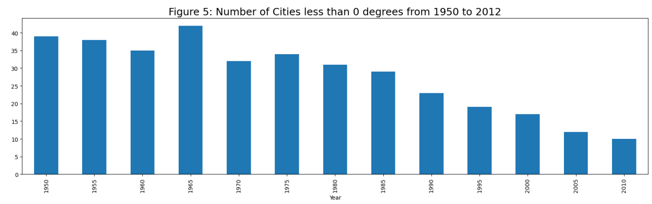


Figure 5: Number of Cities less than 0 degrees from 1950 to 2012

We also presented a histogram displaying the percentage change in annual temperature from 1950 to 2012 across multiple cities in Figure 6. Our analysis indicates that the majority of cities experienced a positive temperature change, with most cities having a percentage change of around 18.5. The overall shape of the histogram closely approximates a normal distribution. Interestingly, we found that from 1950 to 1978, when economic development in China was slow and most areas were rural, most cities had an annual percentage change in temperature of 18.1. However, from 1979 to 2012, during a period of rapid economic development and urbanization, most cities had an annual percentage change in temperature ranging from 18.25 to 19.00. This indicates that temperatures have increased in China over the past few decades, with evidence pointing to urbanization as a significant factor in this trend. While urbanization has brought benefits to humanity in terms of improved living standards and quality of life, it has also led to global warming. The observed increase in temperatures highlights the cost of urbanization and emphasizes the importance of implementing sustainable urban development strategies that mitigate the impact of global warming.

Figure 9, 10, 11 show the geographic distribution of average temperature by province in China. From the Figure 9 and 10 depicting the average temperature of China by provinces and cities in 1950 and 2012, it is difficult to discern any noticeable changes due to the

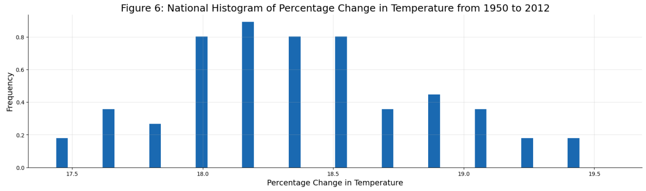


Figure 6: National Histogram of Percentage Change in Temperature from 1950 to 2012

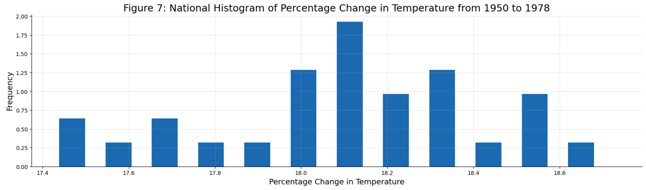


Figure 7: National Histogram of Percentage Change in Temperature from 1950 to 1978

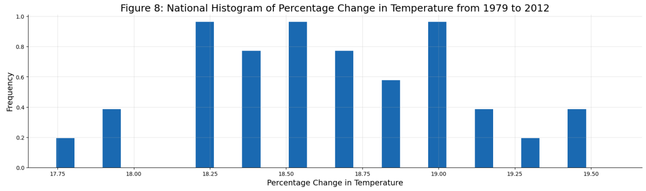


Figure 8: National Histogram of Percentage Change in Temperature from 1979 to 2012

scale used. The difference in temperature from 1950 to 2012 cannot be observed from these maps. But these two graphs still give us some insights of how the temperature differ from south of China with north of China.

The Figure 10 provides a clear and direct visualization of how the average temperature has changed during this time period. We can first see that the overall temperature change is positive, meaning that most of the regions in China is becoming warmer. Additionally, the results are both interesting and intriguing, as they reveal that the south of China, typically known for its warm climate, has experienced a decrease in average temperature. Conversely, the north, known for its colder climate, has seen an increase in average temperature. Xinjiang, in particular, experiences the largest temperature increase as a province in northwest China. This finding was not apparent in the earlier graphs, indicating the importance of this visualization in understanding climate changes in China. Additionally, global warming often leads to more extreme temperature days, which may contribute to the observed temperature differences in the south and north. These findings provide valuable insights into how global warming is impacting different regions of China.

Figure 10: Average Temperature in 2012 by Province and Cities

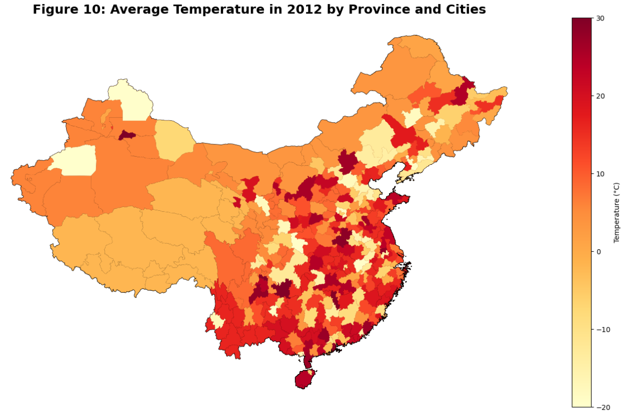


Figure 9: Average Temperature in 1950 by Province and Cities

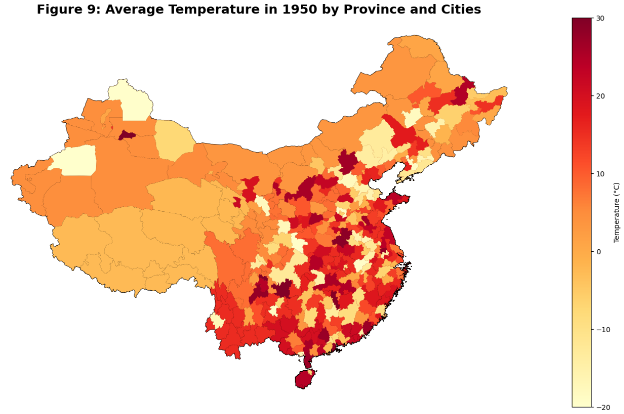
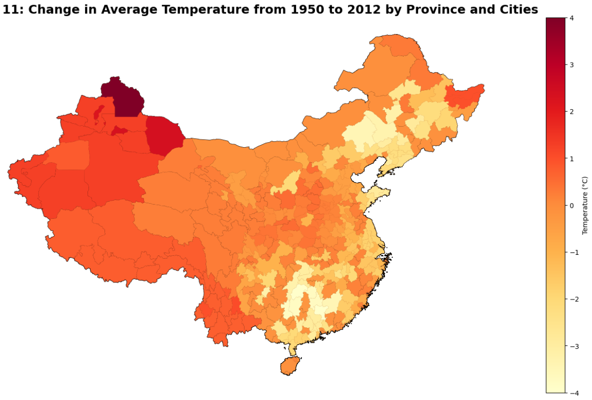


Figure 11: Change in Average Temperature from 1950 to 2012 by Province and Cities



3.2 Effects of Urbanization: Urban and Rural Population

Our research aims to explore the drivers of climate change in China, given that different regions experience different temperature patterns and the average temperature is increasing overall. We will use some data as representative of urbanization to investigate its role in shaping temperature trends in the country. Urban and Rural population is also a direct indicator of urbanization. Unfortunately, our analysis is hindered by missing data for both urban and rural population in China from 1950 to 1959. Simply using a linear regression between years after 1959 and urban or rural population will not work, as it will overpredict the population for 1950 to 1959. An alternative approach would be to use a time series prediction model like the exponential smoothing model. However, applying this model requires a deep understanding and strong application skills that I have not yet acquired.

Therefore, we propose a simpler solution: to use the data of urban and rural population percentage of the total population for 1960 to fill in the missing values from 1950 to 1959. Since the variation in urban and rural population during this time is negligible, we can ignore the changes in population given the broader time range from 1950 to 2012. Furthermore, this period of time (1950-1978) is characterized by slow economic growth and urbanization in China, making it a suitable candidate for this approach. By filling in the missing values with the data from 1960, we can continue our analysis and explore the relationship between urbanization and temperature increase in China.

The Figure 12 and 13 below show the percentage of the total population that lives in urban and rural areas over time in China from 1950 to 2012. These graphs provides valuable insights into the changing demographic patterns in China and the shift of population from rural to urban areas.

We have observed that the percentage of people living in the urban area is increasing over time, while the percentage of people living in the rural area is decreasing over time. This trend is consistent with China's rapid urbanization, which has been driven by economic development and rural-urban migration. The graphs reveal that during the period of 1950 to 1980, there was little change in the percentage of people living in urban areas, and even a slight increase in the percentage of people living in rural areas. This period coincides with the Chairman Mao Zedong era, characterized by slow economic and social development. In contrast, the percentage of people living in urban areas has increased significantly since the 1980s, with a corresponding drop in the percentage of people living in rural areas. This period is marked by the Chairman Deng Xiaoping era, characterized by high-speed industrialization and social development. In recent years, China has experienced significant economic growth, which has led to the rapid expansion of cities and the migration of rural residents to urban areas in search of better job opportunities and higher living standards.

In conclusion, the bar plot and line plot provide valuable insights into the changing demographic patterns in China and give us some insights of how this could relate to the change in temperature in China. Therefore, this shift in population patterns is a important factor to consider when exploring the relationship between climate change and human activities, as urbanization and industrialization contribute significantly to greenhouse gas emissions, leading to climate change.

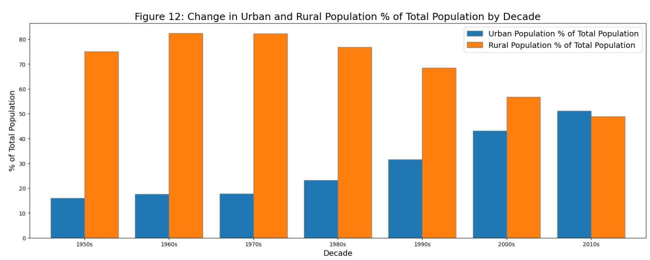


Figure 12: Change in Urban and Rural Population % of Total Population by Decade

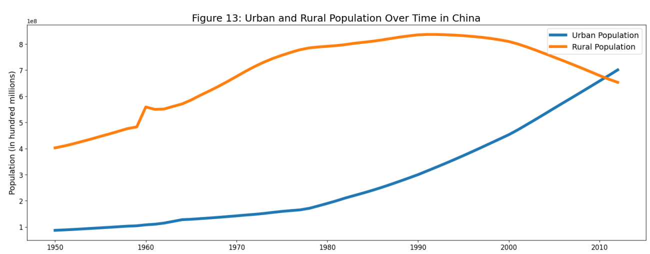


Figure 13: Urban and Rural Population Over Time in China

The scatter plots reveal a significant change in the relationship between urban population and average temperature in China from 1950 to 2012. The Figure 14 (1950-1978) shows no correlation between urban population and average temperature, indicating that during this time when urbanization was slow, the average annual temperature remained relatively stable.

Conversely, the Figure 15 (1978-2012) shows a positive correlation between urban population and average temperature, suggesting that as urbanization accelerated, the average temperature also increased.

Furthermore, analyzing the last two graphs, we observe a positive correlation between urban population and average temperature, as well as a positive relationship between the percentage of the total population residing in urban areas and average temperature in China over the entire period. This means that as the urban population and urbanization increases, the average temperature also tends to increase. This correlation implies that the rising population may affect energy use and fossil fuel consumption, potentially driving climate change. We will explore the relationship between energy use and climate change in project three. This answers the research question in general and suggests that urbanization plays a important role in shaping China as a warming country.

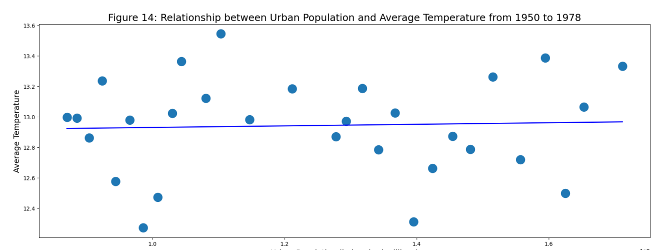


Figure 14: Relationship between Urban Population and Average Temperature from 1950 to 1978

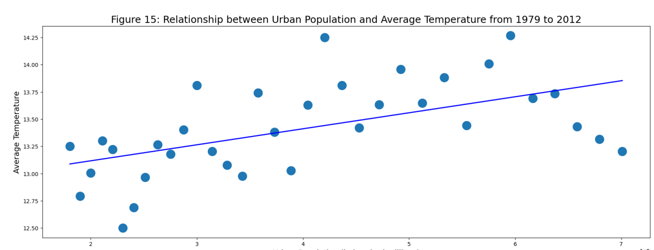


Figure 15: Relationship between Urban Population and Average Temperature from 1979 to 2012

Figure 16: Relationship between Urban Population and Average Temperature from 1950 to 2012

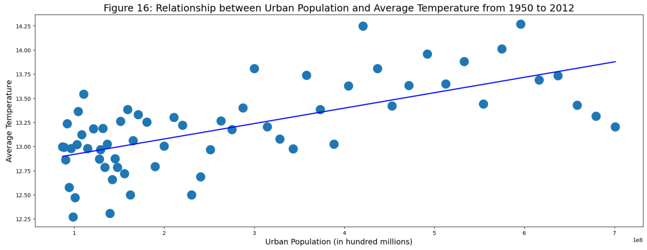
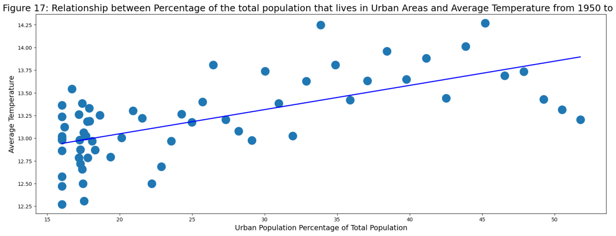


Figure 17: Relationship between Percentage of the total population that lives in Urban Areas and Average Temperature from 1950 to 1978



3.3 Effects of Urbanization: GDP

The first two maps show the GDP of different provinces in China in 1952 and 2012. We chose 1952 instead of 1950 to show the GDP because of missing data from the early years of the People's Republic of China. However, even in 1952, we do not have GDP data for Hainan province because it was part of Guangdong from 1950 to 1988. Despite this missing data, our research and conclusions remain unaffected.

From the Figure 18, we can observe that the average GDP in 1952 by province in China was relatively low, ranging from 1.32 billion yuan in Anhui to 48.41 billion yuan in Zhejiang. We also note that the southern part of China was relatively wealthier than the northern part in 1952.

In the Figure 19, we see that the average GDP in 2012 by province in China was relatively high, ranging from 701.03 billion yuan in Anhui to 57067.92 billion yuan in Zhejiang. Anhui remained the poorest province, while Zhejiang continued to be the richest province in China. Once again, we observe the trend that the southern part of China was wealthier than the northern part. These findings may suggest a relationship between GDP and average temperature, but we will explore this relationship more directly in the next map.

The Figure 20 displaying the difference in GDP between 1952 to 2012 by province in China is highly relevant to our research question, as economic growth is a key driver of urbanization, which, in turn, can impact average temperature trends in a warming world. The similarity between the map of the GDP in 2012 by provinces and the map of the difference in GDP between 1952 and 2012 may be due to the relatively small difference in GDP over time. This is because the GDP values for each province have only ranged from 1.32 billion yuan to 48.41 billion yuan since 1950, which means that the differences between the GDP values for each province may not be very pronounced. As we previously mentioned, the map shows that the south part of China has experienced higher economic growth than the north part of China, which may suggest that the south has undergone more urbanization and subsequently experienced different temperature trends than the north.

Urbanization can impact temperature trends through various mechanisms, such as the urban heat island effect, where urban areas are warmer than surrounding rural areas. However, this pattern is not seen in China. Instead, we observe a clear pattern of decreasing average temperature from 1950 to 2012 in the south part of China, while the north part of China shows a clear pattern of increasing average temperature during the same period.

This finding suggests that the relationship between urbanization and temperature trends may not be straightforward and requires further investigation. We need more evidence to confirm this conclusion and shed more light on the complex interplay between urbanization, economic growth, and temperature trends in China and other regions.

Figure 18: GDP in 1952 by Province in China

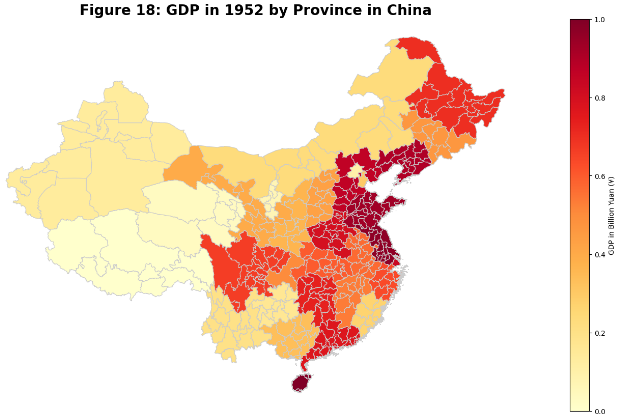
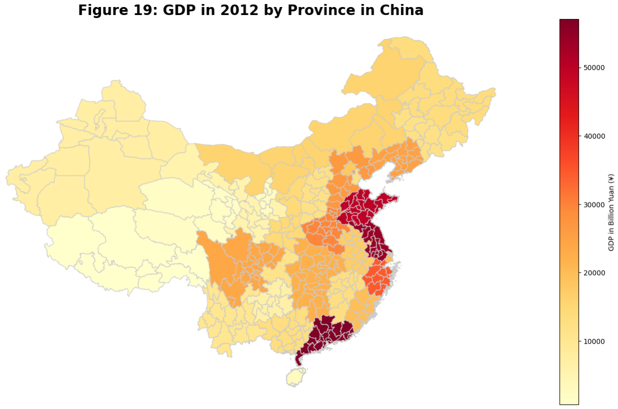


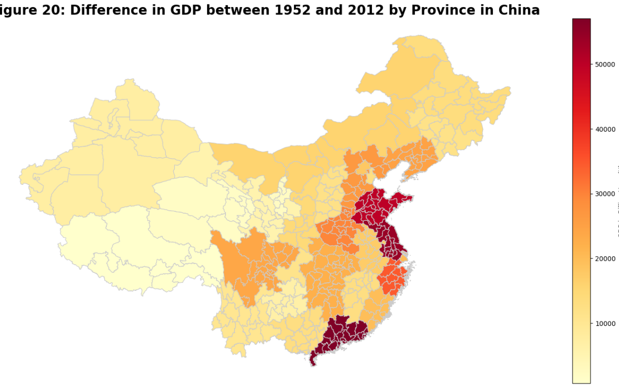
Figure 19: GDP in 2012 by Province in China



3.4 Effects of Urbanization: Urban and Rural Population

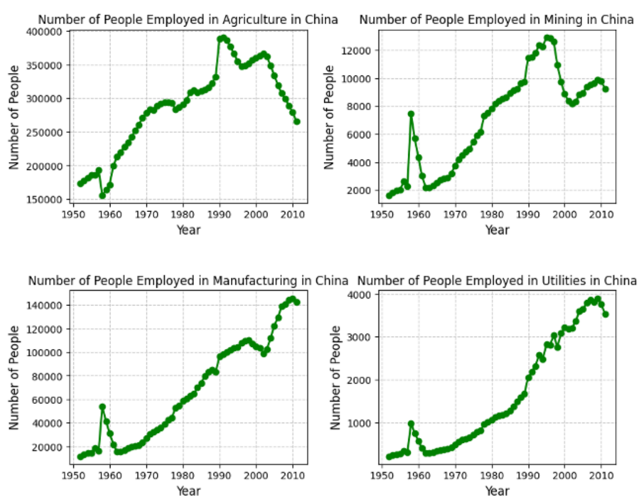
We will begin our analysis by using line plots to depict the employment trends in various industries in China from 1952 to 2011. The first graph displays the number of people employed in agriculture, which accounts for the largest proportion of the workforce. It is evident that the trend was consistently increasing until 1990, after which there was a sharp decline. This can be attributed to the process of urbanization and economic development, which enticed many young people to move from rural to urban areas in search of better opportunities. The combined analysis of these trends in conjunction with other line plots will provide a better understanding of the workforce allocation across different sectors.

Figure 20: Difference in GDP between 1952 and 2012 by Province in China



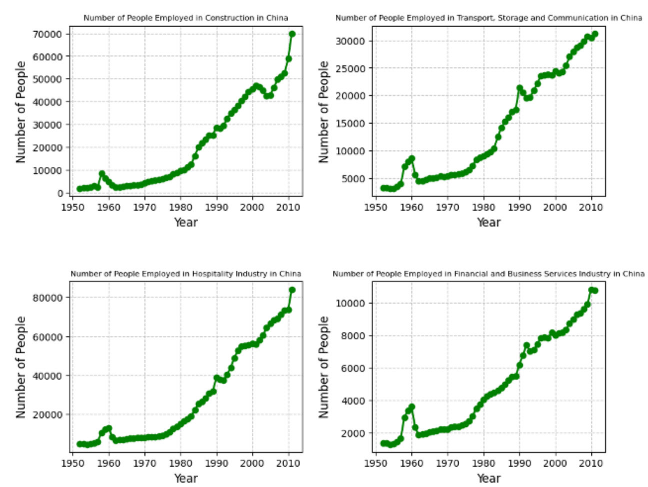
The second graph illustrates the number of people employed in mining in China, which displays a similar trend as that seen in the agriculture sector, with a constant upward trend before 1995 and a sudden drop afterwards. The reasons for this decline are again attributable to urbanization and the emergence of more promising prospects in cities.

The third and fourth graphs showcase the number of People employed in manufacturing and utilities in China, with the former being a significant contributor to the country's economy. The trend in this sector saw a sudden increase in employment in 2000, with the rise of industrialization and factories operating in China. This trend also represents China's shift from an agricultural-based economy to a manufacturing-based one. It is noteworthy that the maximum number of people employed in these sectors was observed in 2010, after which there was a decline. These trends, when analyzed in conjunction with the following four graphs depicting people working in the service industry, provide an insight into China's economic journey.



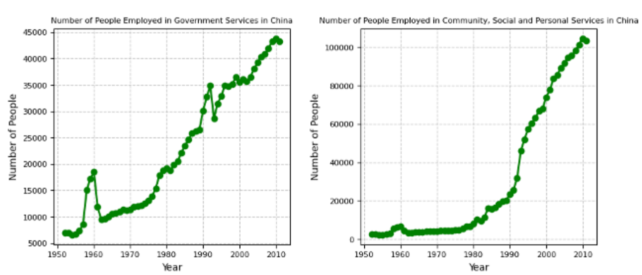
The four graphs depict a notable shift in the employment trends in China, particularly towards service-based industries. The significant increase in the number of people working in the construction industry is an indicator of the need for building infrastructure and buildings to support the process of urbanization in China. This shift towards urbanization has been a significant driving force in shaping the country's economic development and has played a significant role in influencing the temperature trends in the region.

Additionally, the upward trend in the number of people working in the service industry, particularly in trade, restaurants and hotels, transport, storage and communication, and financial and business services sectors, indicates a significant shift in the country's economic priorities from manufacturing to services. This shift is consistent with the broader global trend towards service-based economies, and it highlights China's efforts to diversify its economy and reduce its dependence on the manufacturing sector. These trends suggest that China's urbanization and shift towards a service-based economy have significant implications for the country's economic growth and development.



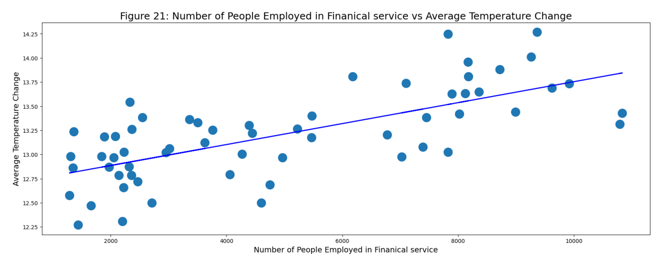
In conclusion, by observing the trend in employment in different industries in China, it is clear that the process of urbanization has played a significant role in shaping the workforce and the economy. As the process of urbanization accelerates, more people move from rural areas to urban centers, leading to a decline in the number of people employed in the agricultural and mining sectors. These trends indicate a shift towards a more industrialized economy and a transition from an agricultural-based to manufacture-based economy, then to a service-based economy.

Moreover, the sudden increase in the number of people employed in the manufacturing sector in China around 2000 also indicates a significant shift in the economy. The development of factories and the growth of the manufacturing sector have been a key driver of the Chinese economy in recent decades. However, the decline in the number of people employed in manufacturing and utilities after 2010 may suggest a new shift in the economy, where services may become more dominant. In conclusion, the line plot suggests that the shift towards an industrialized and service-based economy has been driven in part by the process of urbanization, and that this shift may have implications for temperature trends in the future.



Based on the scatter plots of the number of people working in the financial service industry and trade, restaurants, and hotels industries against average temperature change, it appears that there is a positive correlation between the number of people working in these industries and average temperature change. This suggests that as urbanization and economic development occur, resulting in an increase in the number of people working in these industries, there may be a corresponding increase in average temperature.

The financial service industry is associated with the development of the finance and insurance sectors, which could contribute to economic growth and urbanization. The trade, restaurants, and hotels industry is associated with the development of the service sector, which is a major contributor to economic growth and urbanization. These findings suggest that urbanization and economic development play a role in shaping average temperature trends in a warming world. As more people move to cities and contribute to the growth of these industries, there may be an increase in energy consumption and greenhouse gas emissions, contributing to climate change. This highlights the need for sustainable urbanization and economic development strategies that minimize their impact on the environment and mitigate the effects of climate change.



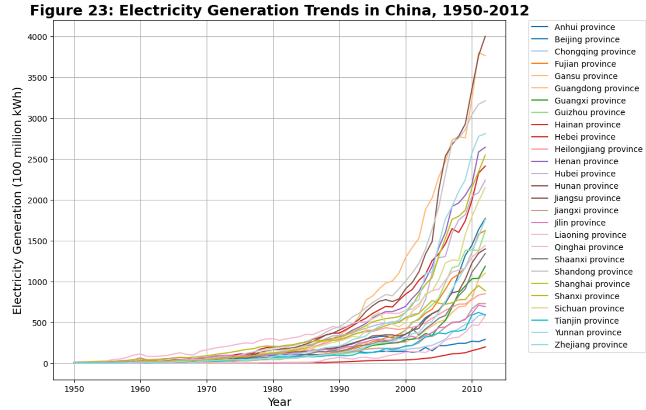
3.5 Effects of Urbanization: Electricity Generation

In this section, we will study the effect of electricity generation on average temperature, and gain a better understanding of the relationship between urbanization, economic development, and average temperature trends in China. This variable also provides valuable insight into the distribution of energy generation across different provinces and how it correlates with economic development. Through creating maps to visualize electricity production across provinces, we can gain a clearer picture of these relationships. Our main indicator years will be 1952 and 2012, and it is important to note that we experienced data loss for some provinces in 1950, such as Zhejiang and Sichuan. We also decided to ignore the observation for Chongqing city since Chongqing was part of Sichuan province until 1997. Additionally, we will exclude Hainan and Tibet since they were founded after 1950 and do not affect our findings.

This line graph (Figure 23) provides valuable insight into the relationship between economic development, electricity generation, and urbanization in China. It illustrates how the increase in electricity generation since the 1980s has coincided with China's rapid economic growth and industrialization, which are closely tied to urbanization. The Northeastern region of China, particularly Liaoning Province, had the highest levels of electricity generation until the 1980s, followed by growth in the Southern region after 1980. The graph also highlights how the increase in electricity generation is likely contributing to rising average temperatures in China, which is a key component of our research question. Therefore, Figure 23 helps to shed light on the potential impact of urbanization and economic development on average temperature trends in China and the importance of considering these factors when examining the impacts of a warming world.

Figure 24 shows the distribution of electricity production across provinces in China in 1952. The darker shades of blue and purple indicate provinces with higher electricity generation, while the lighter

Figure 23: Electricity Generation Trends in China, 1950-2012



shades represent provinces with lower electricity generation. The north part of China, particularly the northeastern region consisting of Jilin, Heilongjiang, and Liaoning provinces, stands out as the area with the highest amount of electricity generation in China. This is likely due to the abundance of natural resources such as coal and oil in this region, which are commonly used to generate electricity. It also indicates that the northeastern region was highly developed at that time with faster pace of industrialization process than other areas. The southern provinces, including Guangdong, Fujian, and Zhejiang, appear to have lower electricity generation compared to the northern provinces, possibly due to differences in industrial development and resource availability.

Figure 25 shows a shift in the distribution of electricity production across provinces in China in 2012 compared to the distribution in 1952. The provinces of Jiangsu, Guangdong, and Shandong had the highest amount of electricity generation, with Jiangsu producing the most at 4001.13 (100 million kWh). This suggests a shift in China's energy landscape towards a more industrialized and urbanized economy, with these provinces being major economic hubs and centers of production. The shift away from the traditional heavy industry in the northeastern provinces towards service-based industries in the southern and eastern provinces may have contributed to this change in electricity production. Additionally, advancements in technology and improvements in energy infrastructure may have also played a role in the increased electricity generation in these provinces.

Figure 26, which shows the difference in electricity production from 1952 to 2012, provides further insight into the changing trends of electricity generation in China over time. We see a similar distribution to the second map, with Jiangsu, Guangdong, and Shandong having the highest electricity production. However, we also see a significant increase in electricity production across most provinces, which highlights the rapid development and urbanization that has occurred in China over the past few decades. This increase in electricity production is likely driven by the growth of urban areas, which require more energy to power buildings, transportation, and infrastructure.

When comparing the changes in electricity production to the changes in temperature across China from 1950 to 2012, we can observe a potential relationship between urbanization and temperature trends. The expansion of urban areas and the subsequent increase in electricity production have resulted in a general increase in temperature in northern parts of China and a decrease in temperature in southern parts of China. This is likely the result of urbanization increasing the frequency of extreme temperature days. The northeastern region of China, in particular, shows a clear trend, as it was the region with high economic development and urbanization, while other parts of China were mostly rural areas. We can see a significant increase in economic development in the northeastern region during 1970, coinciding with a rise in temperature. This may be due to the urban heat island effect, where urban areas are warmer than rural areas due to factors such as increased energy consumption, heat retention by buildings and infrastructure, and reduced green spaces.

Figure 24: Electricity Generation in China, 1952

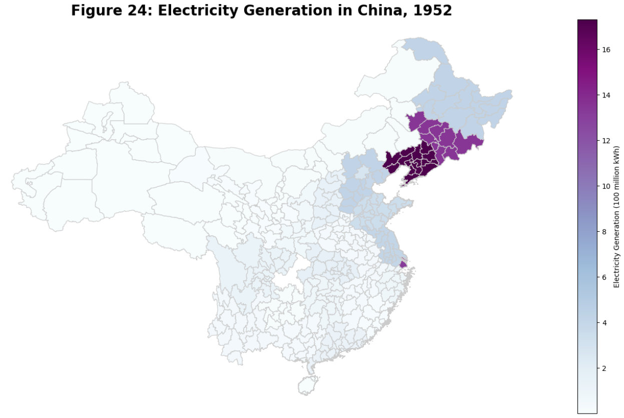


Figure 25: Electricity Generation in China, 2012

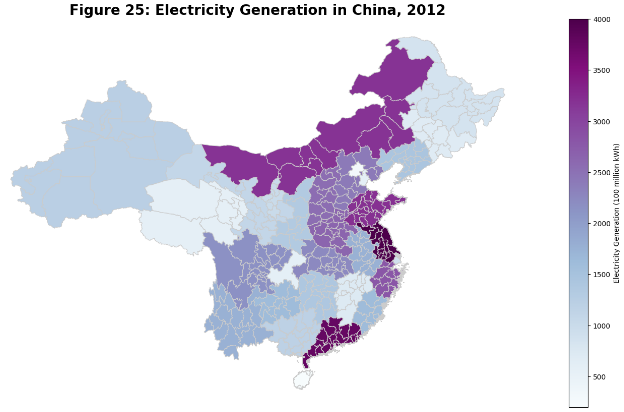
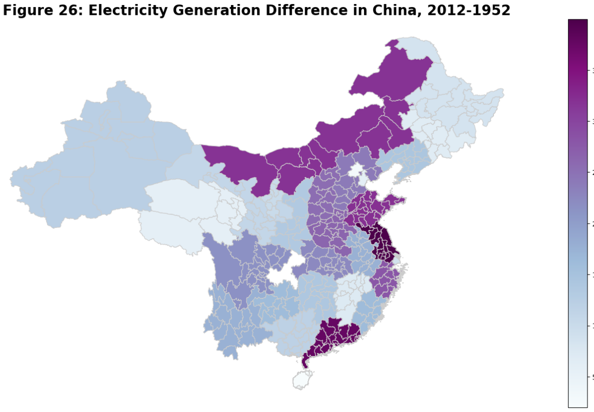


Figure 26: Electricity Generation Difference in China, 2012-1952



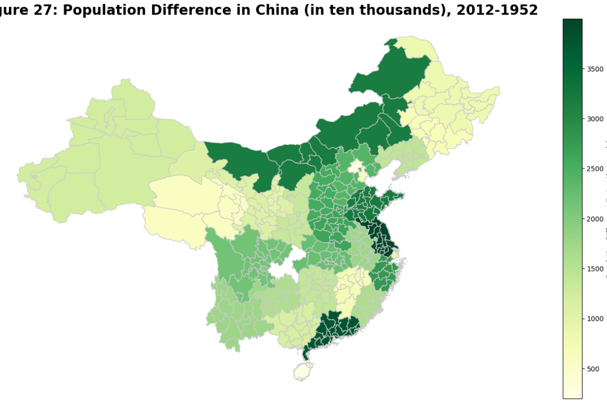
3.6 Effects of Urbanization: Population by province

When considering the factors that may impact average temperature in China, it is important to take into account the significant population boom that has occurred in recent years, particularly in the south region of the country. This is evident when examining the population map by province, which shows a clear trend of increased population in provinces such as Guangdong and Jiangsu, with a population increase of 76,830,000. This population growth is also strongly correlated with the GDP growth in the same regions, as seen in project two.

By including population in our regression models, we can better understand how this growth in population may impact average temperature in different regions of the country. For example, we observe a trend of increasing temperature in the north region of China and decreasing temperature in the south region. This trend may be due to factors such as more extreme hot days in the north and more cold days in the south, which could be influenced by population growth and other factors such as urbanization and industrialization.

Therefore, including population in our regression models can help us to better understand the complex interplay between demographic, economic, and environmental factors that influence average temperature in China. This will enable us to make more accurate predictions about future temperature trends and inform policy decisions related to climate change mitigation and adaptation.

Figure 27: Population Difference in China (in ten thousands), 2012-1952



4 REGRESSION RESULTS

Due to the unique nature of our research topic, collecting over 1000 observations has proven to be a challenge. As we are focusing on the time period between 1950 and 2012 in China, data availability is limited, particularly in the earlier years. In fact, national data was the only available data until 1980, as provincial or city data was not yet recorded. In previous projects, we found that using only national data resulted in less than 100 observations, which did not provide a strong foundation for our research findings.

To address the limitations of our previous regression analysis, we have decided to incorporate additional datasets from the China

Statistical Yearbook. In addition to the GDP and electricity generation datasets used in previous regressions, we will now include data of highway mileage, the number of vehicles, and population by province. We have previously plotted maps for GDP and electricity generation by province, which have shown a linear relationship with average temperature. Specifically, the GDP map has revealed that the south part of China has experienced higher economic growth than the north part of China, suggesting that the south has undergone more urbanization and experienced different temperature trends than the north.

Our analysis has revealed that the expansion of urban areas and the subsequent increase in electricity production have had a significant impact on temperature trends in China. Specifically, there has been a general increase in temperature in northern regions and a decrease in temperature in southern regions. We believe this may be due to urbanization increasing the frequency of extreme temperature days. Additionally, we anticipate that our new datasets of highway mileage, number of vehicles, and population by province are also likely to have a positive relationship with average temperature. This is because as China's economy grows and more cities emerge, the total highway mileage and number of vehicles are expected to increase as well. We have observed similar trends in developed countries like the United States, where large amounts of automobiles and roads were produced. In addition, as we discussed earlier on the effect of urban and rural population on average temperature, we observe a positive correlation between urban population and average temperature, as well as a positive relationship between the percentage of the total population residing in urban areas and average temperature in China over the entire period. These factors are indicative of the level of economic development and can be used as indicators for urbanization, providing a stronger foundation for our regression analysis.

To further investigate these relationships, we plan to conduct data visualization on our newly incorporated datasets. By analyzing the data, we hope to gain a better understanding of the relationship between our independent variables and average temperature.

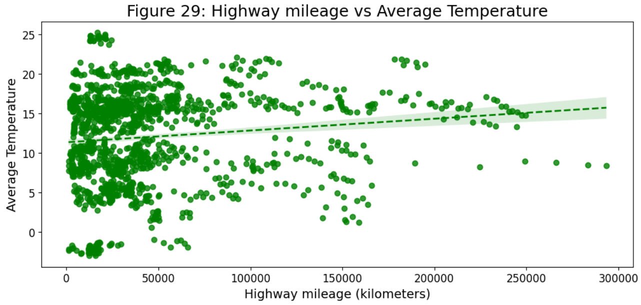
4.1 Non-linearity of independent variables

We will further explore the relationship between each independent variable and the average temperature by creating scatter plots to determine if the relationship is linear or non-linear. This will help us choose which statistical model to use in the subsequent regression analysis.

After examining the scatter plot, we observe that population exhibit a non-linear relationship with the average temperature. Nonetheless, we also notice that population encounter weak heteroskedasticity. To verify the appropriateness of the model, we choose to perform multiple regression initially. Later, we decide to apply a log transformation to these variables in the subsequent multiple regression. By performing multiple regression and applying a log transformation to the population variable, we can better account for the non-linear relationship and assess its impact on the dependent variable.

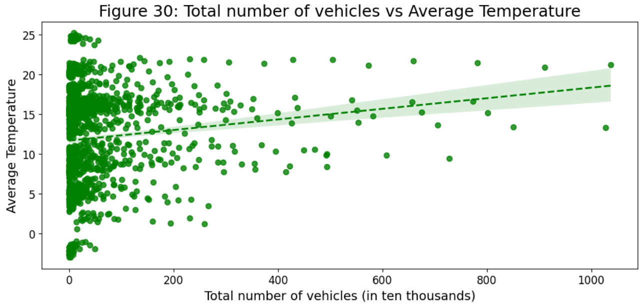
Highway mileage is an important indicator of urbanization and is often associated with economic development and increased population mobility. However, our examination of the scatter plot shows

Figure 29: Highway mileage vs Average Temperature



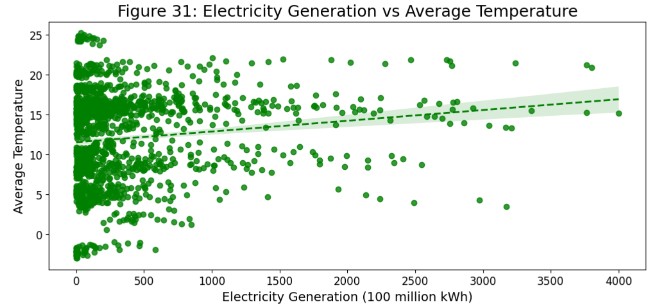
that the relationship between highway mileage and average temperature is non-linear, suggesting that the effect of highway mileage on temperature is not constant across the range of values. Furthermore, we notice that there is strong heteroskedasticity in the relationship, meaning that the variance of the residuals changes as the values of the independent variable increase. In this case, by applying a log transformation to the highway mileage variable, we may be able to better capture its relationship with the average temperature. Additionally, by addressing the heteroskedasticity issue, we can increase the accuracy of our model's predictions and make our results more reliable.

Figure 30: Total number of vehicles vs Average Temperature



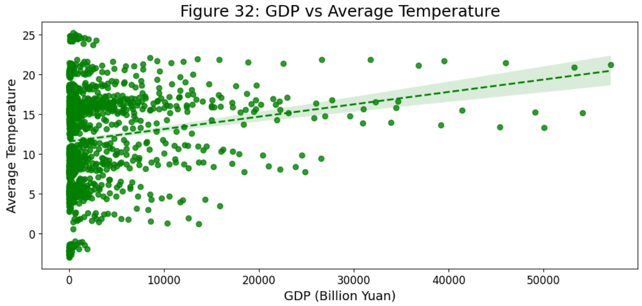
The number of vehicles held in China is a crucial indicator of urbanization, as it reflects the level of economic growth and infrastructure development in a region. However, our examination of the scatter plot shows that the relationship between the number of vehicles and average temperature is non-linear, suggesting that the effect of the number of vehicles on temperature is not constant across the range of values. Furthermore, we notice that there is strong heteroskedasticity in the relationship, meaning that the variance of the residuals changes as the values of the independent variable increase. In this case, by applying a log transformation to the number of vehicles variable, we may be able to better capture its relationship with the average temperature. Additionally, by addressing the heteroskedasticity issue, we can increase the accuracy of our model's predictions and make our results more reliable.

Figure 31: Electricity Generation vs Average Temperature



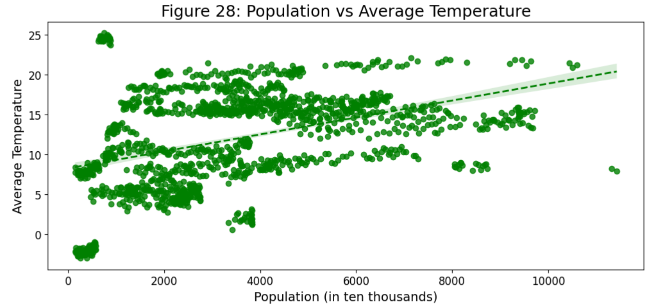
We utilized line plots and maps to visualize the electricity generation by province over time, gaining insight into how electricity production has changed. Electricity generation can potentially explain the variation in average temperature as it is a major source of greenhouse gas emissions, which contribute to climate change. In addition, electricity generation is often linked with industrialization and economic development, both of which can have significant impacts on the environment and temperature patterns. Upon examining the scatter plot, we observed a non-linear relationship between electricity generation and the average temperature. However, we also noted that electricity generation exhibits strong heteroskedasticity. To verify the appropriateness of the model, we choose to perform multiple regression initially. Later, we decide to apply a log transformation to these variables in the subsequent multiple regression. By performing multiple regression and applying a log transformation to the electricity generation variable, we can better account for the non-linear relationship and assess its impact on the average temperature.

Figure 32: GDP vs Average Temperature



We also plotted the map showing the GDP by province in project 2. GDP is an important economic indicator that reflects the overall economic activity. As such, it can be expected to have an impact on various aspects of the region or country, including the temperature. Upon examining the scatter plot, we observed a non-linear relationship between GDP and the average temperature. However, we also noted that GDP exhibits strong heteroskedasticity. To verify the appropriateness of the model, we choose to perform multiple regression initially. Later, we decide to apply a log transformation to these variables in the subsequent multiple regression. By performing multiple regression and applying a log transformation to the GDP variable, we can better account for the non-linear relationship and assess its impact on the average temperature.

Figure 28: Population vs Average Temperature



4.2 Applying log transformation to independent variables

The scatter plots of each independent variable against the dependent variable, average temperature, showed that the relationship between each variable and the dependent variable was non-linear. In order to capture this non-linearity and make our regression models more accurate, we decided to apply a log transformation to each independent variable. After applying the log transformation, we re-examined the scatter plots and observed that the relationships between the independent variables and the dependent variable were now linear. We also noticed that the heteroskedasticity issue observed earlier was resolved. This transformation allowed us to create a more accurate multiple regression model and make reliable predictions about the relationship between the independent variables and the dependent variable.

Figure 33: Total number of vehicles_log vs Average Temperature

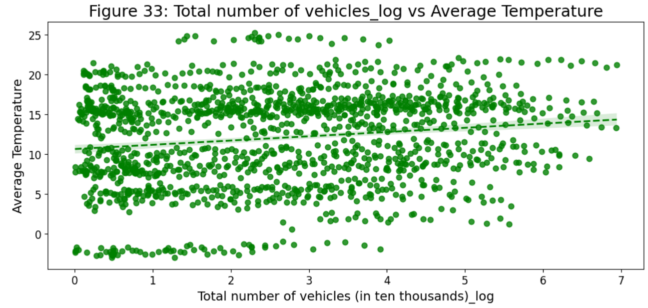


Figure 34: Electricity Generation_log vs Average Temperature

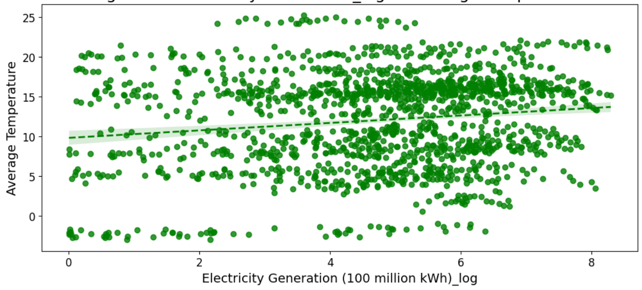


Figure 35: Population_log vs Average Temperature

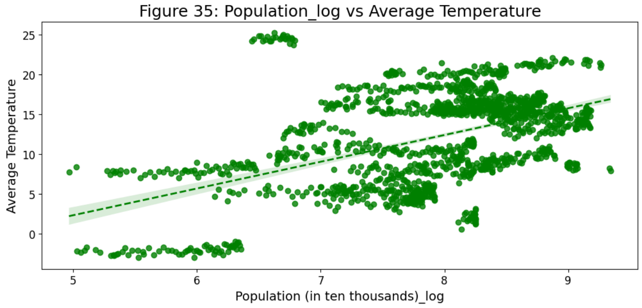


Figure 36: GDP_log vs Average Temperature

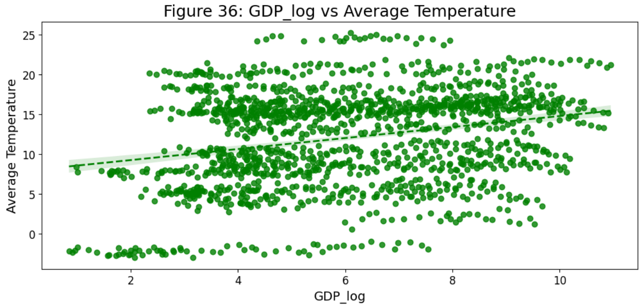
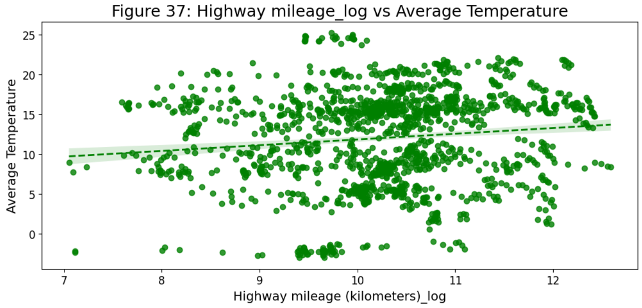


Figure 37: Highway mileage_log vs Average Temperature



4.3 Multiple Linear Regression

Model 1: $\text{Average\_Temperature} = \beta_0 + \beta_1 \text{Population} + \epsilon$

Linear regression may not be sufficient to capture all the changes in average temperature using only one independent variable. Therefore, we opted to use multiple regression models. However, we observed a high R-squared value in the previous simple regression model using population as the independent variable. As such, we decided to incorporate it into our table as an indicator of how our multiple regression is performing. The first model uses population as the independent variable. The second model adds more independent variables, including electricity generation, GDP, total number of vehicles, highway mileage, and population. The third model is based on the second model but adds a dummy variable indicating whether the region is in the south or north part of China. Comparing model 1 to models 2 and 3 clearly demonstrates that the latter provides a more accurate regression outcome.

Model 2:

$$\text{Average\_Temperature} = \beta_0 + \beta_1 \text{Electricity\_Generation} + \beta_2 \text{GDP} + \beta_3 \text{Highway\_mileage} + \beta_4 \text{Population} + \beta_5 \text{Number\_Vehicles} + \epsilon$$

In model 2, the R-squared value indicates that approximately 20.85 percent of the variation in average temperature can be explained by the independent variables, which include electricity generation, GDP, total number of vehicles, highway mileage, and population.

The F-statistic is 76.64, indicating that the model is statistically significant overall. However, when we compare model 2 with model 3, we can see that adding just one more variable, the dummy variable south, significantly improves the model's ability to explain the variation in average temperature. Thus, we decide to move on and find other useful models to explain the relationship.

Model 3:

$$\text{Average\_Temperature} = \beta_0 + \beta_1\text{Electricity\_Generation} + \beta_2\text{GDP} + \beta_3\text{Highway\_mileage} + \beta_4\text{Population} + \beta_5\text{Number\_Vehicles} + \beta_6\text{South} + \epsilon$$

In model 3, around 40.50 percent of the variation in average temperature can be explained by the independent variables, including additional dummy variable south. The F-statistic is also high, at 164.89, indicating that the model is overall statistically significant. In model 3, we can interpret what the coefficient means. After controlling for electricity generation, GDP, total number of vehicles, highway mileage, and population, the south part of China is on average 4.78 celsius degrees hotter than the north part.

After controlling electricity generation, GDP, total number of vehicles, region of China, and highway mileage constant, we are 99 percent confident that every one hundred thousand increase in population is associated with an increase of 0.014 celsius in the average temperature. The coefficient here is larger than in the simple regression, meaning that as population increases, it relates to a higher average temperature. Holding electricity generation, GDP, total number of vehicles, region of China, and population constant, we are 99 percent confident that every one thousand kilometers increase in highway mileage is associated with a decrease of 0.03 Celsius in the average temperature. The result is completely different from simple regression where the coefficient is positive.

Besides these statistically significant coefficients, we also noticed that the number of vehicles coefficient in model 3 is not statistically significant, and we fail to reject the null hypothesis. In addition, the coefficient of electricity generation and GDP is only significant at 90 percent confidence level. It does not mean that these variables have no linear relationship with the average temperature, but that they contribute nothing to the modeling after allowing for all the other predictors.

			Dependent variable:AverageTemperature
	(1)	(2)	(3)
Electricity Generation (100 million kWh)		-0.00341***	-0.00115*
		(0.00072)	(0.00063)
GDP		0.00066***	0.00016*
		(0.00010)	(0.00009)
Highway mileage (kilometers)		-0.00001**	-0.00003***
		(0.00001)	(0.00000)
Population (in ten thousands)	0.00106***	0.00124***	0.00139***
	(0.00006)	(0.00007)	(0.00006)
South			5.69176***
			(0.25979)
Total number of vehicles (in ten thousands)		-0.02261***	0.00028
		(0.00504)	(0.00449)
const	8.24989***	8.58267***	4.77585***
	(0.26180)	(0.26568)	(0.28861)
Observations	1,461	1,461	1,461
R^2	0.16747	0.20846	0.40491
Adjusted R^2	0.16690	0.20574	0.40246
Residual Std. Error	5.46409 (df=1459)	5.33521 (df=1455)	4.62758 (df=1454)
F Statistic	293.49557*** (df=1; 1459)	76.63712*** (df=5; 1455)	164.89048*** (df=6; 1454)
Note:			*p<0.1; **p<0.05; ***p<0.01

4.4 Multiple Linear Regression with Interaction Terms

Incorporating interaction terms can allow us to capture the combined effect of two or more independent variables on the dependent variable, which may not be adequately captured by a simple additive model. By introducing interaction terms into our regression models, we can better capture the complexity of the relationship between the independent variables and the dependent variable, and potentially improve the accuracy of our models.

In this case, we have already developed a multiple regression model (model 3) that is good at explaining the average temperature, but we believe that incorporating interaction terms can make our regression even more accurate. Therefore, we have developed four new models and added four interaction terms to capture the combined effect of certain independent variables on the dependent variable. By doing so, we can potentially improve the explanatory power of our models and obtain more reliable predictions of the average temperature.

Model 4:

$$\text{Average\_Temperature} = \beta_0 + \beta_1\text{Electricity\_Generation}+$$
$$\beta_2\text{GDP} + \beta_3\text{Highway\_mileage} + \beta_4\text{Population}+$$
$$\beta_5\text{Number\_Vehicles} + \beta_6\text{South}+$$
$$\beta_7\text{South} \times \text{Electricity\_Generation} + \epsilon$$

The fourth model adds the interaction term 'Electricity South', which captures the potential nonlinear effects of electricity generation on temperature in different regions. However, the coefficient of 0.00020 is not significant, and we cannot provide an explanation for the variable. Thus, this model is not helpful and we choose to move on.

Model 5:

$$\text{Average\_Temperature} = \beta_0 + \beta_1\text{Electricity\_Generation}+$$
$$\beta_2\text{GDP} + \beta_3\text{Highway\_mileage} + \beta_4\text{Population}+$$
$$\beta_5\text{Number\_Vehicles} + \beta_6\text{South} + \beta_7\text{South} \times \text{Population} + \epsilon$$

The fifth model adds the interaction term 'Population South'. Unlike the first model, the coefficient is now significant, and we can provide an economic interpretation for it. Holding other factors constant, for every one million increase in population in the south region of China, we are 99 percent confident that it will cause the average temperature to increase around 0.05 Celsius more than it would in the north areas. In other words, the effect of population on temperature is stronger in the south region compared to the north region.

Model 6:

$$\text{Average\_Temperature} = \beta_0 + \beta_1\text{Electricity\_Generation}+$$
$$\beta_2\text{GDP} + \beta_3\text{Highway\_mileage} + \beta_4\text{Population}+$$
$$\beta_5\text{Number\_Vehicles} + \beta_6\text{South} + \beta_7\text{South} \times \text{GDP} + \epsilon$$

Similar to the fourth model, the sixth model does not give us clear information since the interaction term 'GDP South' is not significant, and we cannot provide an economic interpretation for it. Thus, this model is not helpful and we choose to move on.

Model 7:

$$\text{Average\_Temperature} = \beta_0 + \beta_1 \text{Electricity\_Generation} + \beta_2 \text{GDP} + \beta_3 \text{Highway\_mileage} + \beta_4 \text{Population} + \beta_5 \text{Number\_Vehicles} + \beta_6 \text{South} + \beta_7 \text{South} \times \text{Highway\_mileage} + \epsilon$$

The seventh model adds the interaction term 'Highway South', which has a statistically significant coefficient, and we can provide an economic interpretation for it. Holding other factors constant, for every one thousand increase in highway mileage in the south region of China, we are 99 percent confident that it will cause the average temperature to increase around 0.02 Celsius more than it would in the north areas. The effect of the number of highway mileage on temperature is stronger in the south region compared to the north region.

After conducting our analysis, we found that population has a significant impact on the average temperature. Our observations also indicated that the effect of population on temperature is more pronounced in the south region than the north region. This highlights the significance of population in shaping the climate patterns of a region. Therefore, we have decided to incorporate the variable Population South in our future regression models as an independent variable to better understand and account for its impact on average temperature. By including this variable in our analysis, we can improve the accuracy of our model and provide a more comprehensive understanding of the factors that contribute to changes in average temperature.

				Dependent variable:AverageTemperature
	(1)	(2)	(3)	(4)
Electricity Generation (100 million kWh)	-0.00122*	-0.00103	-0.00115*	-0.00074
	(0.00065)	(0.00063)	(0.00063)	(0.00064)
Electricity_South	0.00020			
	(0.00046)			
GDP	0.00015	0.00013	0.00015	0.00012
	(0.00009)	(0.00009)	(0.00010)	(0.00009)
GDP_South			0.00001	
			(0.00004)	
Highway mileage (kilometers)	-0.00003***	-0.00003***	-0.00003***	-0.00004***
	(0.00000)	(0.00000)	(0.00000)	(0.00001)
Highway_South				0.00002***
				(0.00001)
Population (in ten thousands)	0.00139***	0.00113***	0.00139***	0.00139***
	(0.00006)	(0.00008)	(0.00006)	(0.00006)
Population_South		0.00051***		
		(0.00011)		
South	5.62404***	3.88016***	5.66253***	4.95736***
	(0.30231)	(0.45594)	(0.28113)	(0.34719)
Total number of vehicles (in ten thousands)	0.00071	0.00143	0.00068	0.00153
	(0.00460)	(0.00447)	(0.00473)	(0.00450)
const	4.81674***	5.83275***	4.79209***	5.26485***
	(0.30338)	(0.36079)	(0.29479)	(0.32632)
Observations	1,461	1,461	1,461	1,461
R^2	0.40499	0.41427	0.40494	0.40902
Adjusted R^2	0.40213	0.41145	0.40208	0.40617
Residual Std. Error	4.62887 (df=1453)	4.59264 (df=1453)	4.62906 (df=1453)	4.61319 (df=1453)
F Statistic	141.28362*** (df=7; 1453)	146.80899*** (df=7; 1453)	141.25533*** (df=7; 1453)	143.65906*** (df=7; 1453)
Note:				*p<0.1; **p<0.05; ***p<0.01

4.5 Multiple Linear Regression with Log terms

Model 8:

$$\text{Average\_Temperature} = \beta_0 + \beta_1 \log(\text{Electricity\_Generation}) + \beta_2 \log(\text{GDP}) + \beta_3 \log(\text{Highway\_mileage}) + \beta_4 \log(\text{Population}) + \beta_5 \log(\text{Number\_Vehicles}) + \beta_6 \text{South} + \epsilon$$

In the model 8, we have applied a log transformation to each independent variable to account for non-linearity. The high F-statistic value of 351.20 suggests that the model is overall statistically significant. Using the logged version of independent variables to explain average temperature has yielded a more accurate and reliable model, as evidenced by the high R-squared value of 0.5917, indicating that approximately 59.17 percent of the variation in average temperature can be explained by the variation in independent variables.

$$\textbf{Model 9:}\ \text{Average\_Temperature} = \beta_0 + \beta_1 \log(\text{Electricity\_Generation}) + \beta_2 \log(\text{GDP}) + \beta_3 \log(\text{Highway\_mileage}) + \beta_4 \log(\text{Population}) + \beta_5 \log(\text{Number\_Vehicles}) + \beta_6 \text{South} + \beta_7 \text{South} \times \text{Population} + \epsilon$$

Model 9, which includes an additional independent interaction term, exhibits the highest R-squared and adjusted R-squared values among all the models we examined, with a slight decrease in the F statistic. However, we cannot conclude that the second model is less effective in explaining average temperature. An economic interpretation of the coefficients indicates that all independent variables, except for the logged total number of vehicles, are statistically significant. The negative coefficient of electricity generation (-0.764) suggests that, holding other factors constant, a 10 percent increase in electricity generation is associated with a decrease of approximately 0.08 degrees Celsius in average temperature. This finding is contrary to our initial expectation, where we hypothesized a positive relationship between energy generation and average temperature. However, we cannot completely rule out the possibility that electricity generation contributes to warming China. To gain further insight, we conducted regression tree analysis in the machine learning section. Additionally, a 10 percent increase in GDP is associated with a increase of approximately 0.15 degrees Celsius in average temperature, indicating that economic development and urbanization contribute to the warming process in China. Moreover, a 10 percent increase in highway mileage is associated with a decrease of approximately 0.29 degrees Celsius in average temperature, while a 10 percent increase in population is associated with an increase of approximately 0.52 degrees Celsius in average temperature. The positive effect of population on average temperature is the most prominent and significant among all the independent variables.

Further investigation of the interaction term reveals that, holding other factors constant, a one million increase in population in the south region of China is associated with a decrease of approximately 0.032 degrees Celsius more in average temperature than in the north region, with 99 percent confidence. This suggests that the effect of population on temperature is weaker in the south region compared to the north region, which contradicts our previous regression outcome. However, when we combine this new result with our previous findings, which indicate that the north region experiences a higher temperature change than the south region of China, a more comprehensive picture emerges. As population growth and urbanization continue to accelerate, it generates more extreme temperature days, where the north part of China experiences more high-temperature days, while the south region experiences more low-temperature days. It may be beneficial to further analyze the differences between the north and south regions by examining the results of the regression tree in the machine learning section. This finding sheds light on the complex relationship between population, urbanization, and temperature in different regions of China, highlighting the importance of considering regional variations when studying the impacts of urbanization on climate change. As a result, this well-constructed model provides valuable insights into how urbanization influences the average temperature and contributes to the warming trend in China.

	Dependent variable:AverageTemperature	
	(1)	(2)
Electricity Generation (100 million kWh)_log	-0.81755***	-0.76429***
	(0.15305)	(0.15333)
GDP_log	1.54720***	1.45240***
	(0.24687)	(0.24761)
Highway mileage (kilometers)_log	-2.79808***	-2.87252***
	(0.18020)	(0.18092)
Population (in ten thousands)_log	5.00087***	5.51620***
	(0.18973)	(0.24334)
Population_South		-0.00032***
		(0.00010)
South	5.97075***	7.21089***
	(0.21909)	(0.42845)
Total number of vehicles (in ten thousands)_log	-0.39545	-0.25953
	(0.30049)	(0.30214)
const	-6.83217***	-10.25283***
	(1.40356)	(1.72919)
Observations	1,461	1,461
R^2	0.59171	0.59487
Adjusted R^2	0.59003	0.59292
Residual Std. Error	3.83307 (df=1454)	3.81955 (df=1453)
F Statistic	351.20330*** (df=6; 1454)	304.78388*** (df=7; 1453)
Note:		*p<0.1; **p<0.05; ***p<0.01

5 MACHINE LEARNING

In order to further enhance our analysis, we decided to use regression tree to gain more insights into the relationship between the independent variables and the dependent variable. We started by selecting the independent variables that we included in our ideal regression model, and we fitted the decision tree model using these variables. We set the maximum depth of the tree to 3 to avoid overfitting, and we plotted the resulting tree to visualize the splits and branches.

To evaluate the performance of our decision tree model, we calculated the mean squared error between the predicted values and the actual values. Interestingly, we found that the mse for the predications is 12.09. The decision tree model provides a useful visualization of the data and can identify nonlinear relationships that may be missed in linear regression models.

$$\frac{1}{N}\sum_{i=1}^{N}\big(\text{AverageTemperature}_i - (\beta_0 + \beta_1\log(\text{Electricity Generation}_i)$$
$$+\beta_2\log(\text{GDP}_i) + \beta_3\log(\text{Highway mileage}_i) + \beta_4\log(\text{Population}_i)$$
$$+\beta_5\log(\text{Total number of vehicles}_i) + \beta_6\text{South}_i)\big)^2 \tag{1}$$

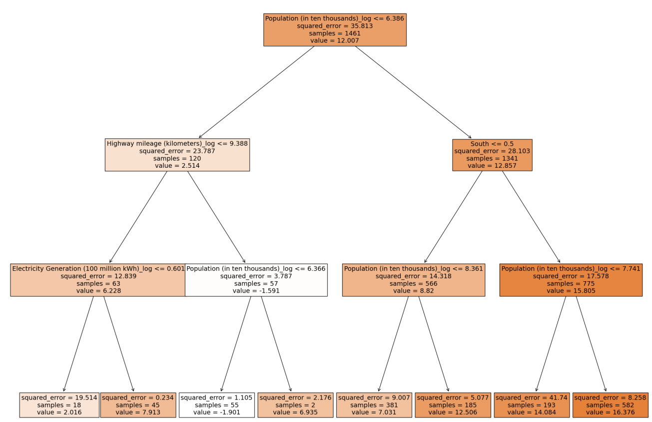
The objective function for a regression tree with the given variables is to minimize the sum of squared differences between the observed average temperature values and the predicted values of average temperature from a regression tree model. The equation uses the average temperature as the dependent variable and includes independent variables such as the logarithms of electricity generation, GDP, highway mileage, population, total number of vehicles, and a dummy variable for South region. The coefficients represent the impact of each independent variable on the average temperature, and the goal of the regression tree is to find the best values for these coefficients that minimize the overall error between the observed and predicted values.

5.1 Regression Tree

When examining the initial node of the regression tree, we observed that for provinces with a population growth rate below 6.386 percent, the predicted average temperature is around 12.007 Celsius degrees. In the second layer of the tree, the data is split based on two independent variables: highway mileage and population. The left branch of the tree shows that for provinces where the highway mileage growth rate is less than or equal to 9.388 percent and population growth rate is below 6.386 percent, the predicted average temperature is 2.514 degrees Celsius. On the other hand, the right branch of the tree indicates that for provinces where the population growth rate is less than or equal to 7.936 percent but greater than 6.386 percent, the predicted average temperature is 12.857 degrees Celsius. The same applies to the third layer.

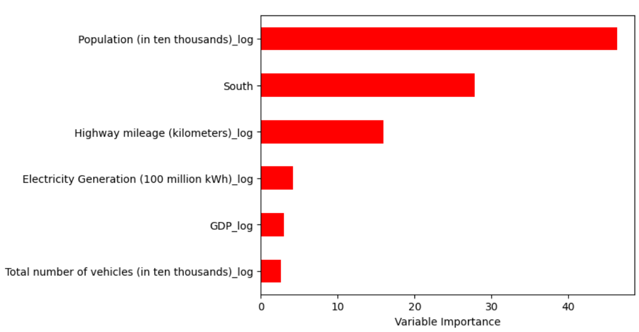
The final layer of the regression tree provides the most insightful and useful information. Our multiple regression analysis showed that an increase in electricity generation is associated with a decrease in average temperature, which contradicts what we proposed earlier, and we could not interpret the relationship before. However, after plotting the regression tree, we gained insight into how electricity generation affects the average temperature in China. The tree showed that for provinces with population growth rates greater than 6.386 percent and highway mileage growth rates less than 9.388 percent, those with electricity generation growth rate higher than 0.6 percent will generally experience a higher temperature of 5.9 degrees Celsius. This result is not what we observed based on multiple regressions, and it supports our hypothesis about the role of urbanization in shaping average temperature trends.

Furthermore, the regression tree reveals that population is a crucial factor in driving the increase of average temperature. For provinces located in the northern region of China, those with a population growth rate exceeding 8.36 percent are expected to experience a temperature rise of approximately 5.5 degrees Celsius. Similarly, in the southern region, provinces with a population growth rate surpassing 7.74 percent will experience a temperature increase of around 2.29 degrees Celsius. These results confirm that provinces with a higher population, regardless of their geographic location, are more prone to experiencing elevated temperatures, thereby reinforcing the findings obtained from our multiple regression analysis. In the multiple regression result from model 9, we get the conclusion that the effect of population on temperature is weaker in the south region compared to the north region, which is justified here. The observed population growth and urbanization may well be contributing to the escalating average temperature in China.



5.2 Random Forest (RF) and Importance Matrix

The results of our Random Forest model provide significant insights into the relationship between urbanization and average temperature in China. The drastic reduction in mse from 12.09 to 0.36 indicates the high accuracy of the model. The importance matrix showed that the population variable has a significant impact on reducing the error, which confirms our previous findings on the role of population growth in shaping average temperature trends. Furthermore, we found that creating a dummy variable to classify regions into north and south was effective in reducing the error, suggesting that regional differences play an important role in this relationship. Finally, our results indicate that highway mileage also contributes to shaping average temperature trends. Our findings provide empirical evidence of the significant impact of urbanization and related factors such as population growth and infrastructure development on average temperature trends, highlighting the importance of taking into account these factors in developing policies to mitigate the impacts of global warming.



5.3 Comparison with Regression Output

The main difference between the results from running a regression and a regression tree is the ability to capture non-linear and interactive effects. In a regression model, we assume a linear relationship between the dependent and independent variables, and we can only estimate the effect of each independent variable on the dependent variable individually, holding all other variables constant. However, in a regression tree, we can capture non-linear relationships and interactions between variables by recursively partitioning the data into subgroups based on the values of the independent variables.

In model 9 of our multiple regression analysis, we observed that there was a negative relationship between electricity generation and average temperature. Specifically, a 10 percent increase in electricity generation was associated with a decrease of approximately 0.08 degrees Celsius in average temperature, which was contrary to our initial hypothesis of a positive relationship between the two variables. However, the output from our regression tree provided further insights into this relationship. We found that provinces with higher electricity generation growth rates, exceeding 0.6 percent, experienced an increase in average temperature of 5.9 degrees Celsius, indicating that there is a positive relationship between the two variables, as we initially hypothesized. The urban heat island effect, where regions that generate more electricity are warmer than those that generate less, could be responsible for this positive relationship due to factors such as increased energy consumption, heat retention by buildings and infrastructure, and reduced green spaces.

Furthermore, the result from the regression tree validates our previous findings from multiple regression. We also found that the effect of population growth on temperature differed between the northern and southern regions of China. While our initial solution was unable to explain this difference, the outcome from our regression tree suggests that provinces with higher population growth rates, regardless of their location, are more prone to experiencing elevated temperatures. In the northern region, provinces with a population growth rate exceeding 8.36 percent were expected to experience a temperature rise of approximately 5.5 degrees Celsius, while in the southern region, provinces with a population growth rate surpassing 7.74 percent would experience a temperature increase of around 2.29 degrees Celsius. Our importance matrix also suggests that the population variable had a significant impact on reducing the error, which confirms our previous findings on the role of population growth in shaping average temperature trends.

6 CONCLUSION

The research found that urbanization plays an important role in shaping average temperature trends in China from 1950 to 2012. As the graphs, multiple regression, and regression tree analyses suggest, we see that the temperature is increasing at a higher speed from 1979 to 2012 (when urbanization happened and the economy starts to bloom in China) compared with 1950 to 1978 (when the economy is growing very slowly). In addition, the average annual temperature is increasing in China from 1950 to 2012. More and more cities experienced extremely high-temperature years and fewer cities experienced low-temperature years. As urban population and the percentage of the urban population increases, the average temperature in China also increases. Furthermore, as China transitioned from an agriculture-based economy to a service-based economy, we also observed an increasing trend in temperature.

Our analysis has shown that population growth and urbanization are likely contributing factors to the rise in average temperature in China. Our regression tree analysis also provided additional insights by highlighting the impact of electricity generation growth rates on temperature changes. These findings underscore the urgent need for policymakers to prioritize sustainable urbanization and environmental conservation efforts to mitigate the negative impact of urbanization on the climate.

In conclusion, our study provides a comprehensive analysis of the impact of urbanization on temperature trends in China, utilizing both multiple regression and regression tree methods. The findings underscore the importance of considering urbanization as a key factor in climate change and highlight the need for policymakers to prioritize sustainable urbanization and environmental conservation efforts. Future research can build upon our findings by exploring additional indicators beyond urban and rural population, such as poverty levels and pollution emissions, to provide a more comprehensive understanding of the impact of urbanization on temperature changes.

REFERENCES

[1] Li, C., Yan, J., & Xu, Z. (2021). How Does New-Type Urbanization Affect the Subjective Well-Being of Urban and Rural Residents? Evidence from 28 Provinces of China. Sustainability.

[2] Li, Y., Zhu, L., Zhao, X., Li, S., & Yan, Y. (2013). Urbanization Impact on Temperature Change in China with Emphasis on Land Cover Change and Human Activity. Journal of Climate, 26, 8765-8780.

[3] Zhang, N., Gao, Z., Wang, X., & Chen, Y. (2010). Modeling the impact of urbanization on the local and regional climate in Yangtze River Delta, China. Theoretical and Applied Climatology, 102, 331-342.

[4] Fang, F., Guo, J., Sun, L., Wang, J., & Wang, X. (2014). The effects of urbanization on temperature trends in different economic periods and geographical environments in northwestern China. Theoretical and Applied Climatology, 116, 227-241.

[5] Sun, Y., Zhang, X., Ren, G., Zwiers, F.W., & Hu, T. (2016). Contribution of urbanization to warming in China. Nature Climate Change, 6, 706-709.

[6] World Bank Group, Climate Change Knowledge Portal.

[7] https://www.kaggle.com/datasets/amritharj/population-of-china-19502022?resource=download

[8] https://doc.wendoc.com/b0b03a03c35395c84a353447d2fc02b48a1224354.html

[9] http://www.stats.gov.cn/english/Statisticaldata/AnnualData/