# The influence of urbanization on climate measurements

Onno Valkering [*]
VU University Amsterdam
Department of Computer Science
onno.valkering@gmail.com

Kashif Zahid [†]
VU University Amsterdam
Department of Computer Science
chkashkhan@gmail.com

## ABSTRACT

This paper considers one of the arguments against global warming and climate change, namely the hypothesis that the rise in temperature is caused by urbanization around weather stations from which data is used used to measure the change in temperatures. Experiments have been conducted to determine the amount of the weather stations in urban areas, and the correlations between the increase of temperature measurements and urbanization of the surrounding area. The results of these experiments show that there are indeed some areas with high correlations, but the amount is practically too low to have any significant impact on the overall picture of global warming. Further work involves the acquisition of more fine-grained information about urban area to provide more specific results.

## 1. INTRODUCTION

The majority of climate change experts agree that humans are the cause for the current global warming [1]. Yet, there still exists a group that believe global warming is a hoax. This paper considers one of the arguments against global warming and climate change, namely the hypothesis that the rise in temperature is caused by urbanization around weather stations from which data is used used to measure the change in temperatures. Although large cities may be be warmer, due to *Urban Heat Islands (UHI)* [2], usually no empirical evidence is provided when the aforementioned argument is used[1] by climate change skeptics.

To gain insight in to which extend climate measurements are influenced by urbanization, we analyzed a large amount of weather station measurements and urbanization indicators (see section 3.1). The analysis of these data sets required *Large Scala Data Engineering* principles and practices due to the vast amount and large sizes. This paper summarizes the approach, results and conclusions of this endeavour. Starting with a discussion of related work in section 2, the research questions in section 3 and the project setup in section 4. Then the results of the data analysis and corresponding experiments are addressed in section 5. We finalize this paper with discussions and conclusions in section 7.

## 2. RELATED WORK

Related to our research of the influence of urbanization on climate change is [6]. This research argues that it is not the growth in population and urbanization that result in climate change, but rather the consumption rate and intensity that is increasing. This results in more emission of greenhouse gas, (GHG) which negatively influences climate change. It counters the assumption that higher urbanization result in higher GHG emissions, by showing examples of cities that have a high urbanization but a relatively lower GHG emission that other, possibly less urbanized areas. This is made possible by making cities more energy efficient. In [5] the phenomenon of UHIs is discussed in detail. This effect causes urban areas to be warmer than the surrounding areas, and is one of the reasons some people doubt the global warming effects.

On the technical side [10] provides us with an inventorization of possible machine learning analytics for historical data. It discusses the pros and cons of several machine learning libraries, including Apache Spark MLlib, Apache Hama and Oryx. Also, [4] is related. It describes a plugin for Apache Spark, "H5Spark", that allows the use of the binary HDF5/NetCDF4 files within the Spark ecosystem. These files are often used by data sets describing geological data.

## 3. RESEARCH QUESTIONS

As discussed in section 1, we want to gain insight in the influence of urbanization on climate measurements. We intent to achieve this based on the following two research questions:

- What proportion of weather stations is located in or close-by urban areas?
- Does a strong correlation exist between the temperature measurements and the urbanization around a weather station?

Let us more specifically explain the research questions. First, since global warming calculations are based on a large amount of worldwide weather stations measurements, it is interesting to determine the amount of weather stations that are in or close-by urban areas. Since these are the weather stations that could potentially influence the overall picture of global warming by the result of urbanization. The next step is then, for each of the weather stations in urban areas, to discover how strong, if any, the correlation between the rising temperature measurements and urbanization is, answering the second research question.

[*] Studentnumber: 2557962
[†] Studentnumber: 2577573
[1] See for example: `http://www.globalclimatescam.com/opinion/top-ten-reasons-climate-change-is-a-hoax`

Specific measures exist to quantify urbanization, such as *Urban Intensity Index (UII)* and *Common Urban Intensity Index (CUII)*. Both of these measurements rely on a vast amount of data about a particular urban area, in addition to land usage, such as infrastructure, population and socioeconomic characteristics [8]. The information required for worldwide coverage of these indexes is not always available publicly and/or in accessible formats, let alone for multi-year time span.

As an alternative, we use three variables that are generally associated with urban areas:

- urban land, measured in km$^2$ covered;
- population, total number of residents;
- population density, measured in residents per km$^2$.

As a threshold, we consider areas only as "urban" if it has an urban land coverage of $\geq$ 10 km$^2$ and a population of $\geq 300,000$ residents in 2014 (based on United Nations qualifications, see section 3.1)

## 3.1 Data sets and time period

We used three data sets in order to answer the research questions, the first two of the data sets described below originate from the US *National Oceanic and Atmospheric Administration (NOAA)*[2] the third data set is published by the *United Nations Department of Economic and Social Affairs (UN DESA)*[3].

We used the weather station measurements from the *Integrated Surface Data (ISD)* data set. This data set contains hourly climate measurements for more than 20,000 weather station worldwide, in the period of 1901 to 2016 [7]. Urbanization data is retrieved from *Historical Database of the Global Environment (HYDE)* contained in the *Historical Land-Cover Change and Land-Use Conversions* data set. This data set contains information about (estimated) global land use for the period 1770-2010 [3]. For population figures we used the *Annual Population of Urban Agglomerations* data set. This data set contains a lineup of worldwide cities with their annual (estimated) population (1950-2030) on condition that they have $\geq 300,000$ residents in 2014 [9].

We based our target time period of the overlap of time periods of the first two data sets, which is 1901 to 2010. However, we corrected this target time period to 1950 to 2010 after investigation of the amount of weather measurements from distinct weather stations. The amount of weather stations becomes of a more considerable size from 1950 onward (see figure 1, the observable interruption around 1970 is caused by the transition of analog to digital reporting of data [7]). Also, before 1950 world regions other than North America and Europe are too under-represented and conveniently the data set with population counts for urban areas starts from the year 1950.

# 4. PROJECT SETUP

In order to process and analyze the data sets with a combined, decompressed, size of $\approx$ 250GB we need an appropriate setup that is capable processing this amount of data. We picked Apache Spark as our main cluster computing framework of choice. Spark provides us with a productive development environment and provides a handy shell for manual exploration of data sets. In addition to the core Spark framework, we used the Spark SQL and Spark MLlib components. Since our data can be represented using tabular formats, we used Spark SQL to query the data in a SQL-fashion. The MLlib component has been used to perform correlation tests. Also, we used an unofficial Spark plugin called Magellan[4] that extends Spark SQL with geospatial operations. Besides a local installation of Spark we had access to the SURFsara Hadoop cluster to run our Spark applications (written in Java and Scala). In addition to Spark we also used a great amount of Python and R scripting to locally explore, alter and/or visualize data.

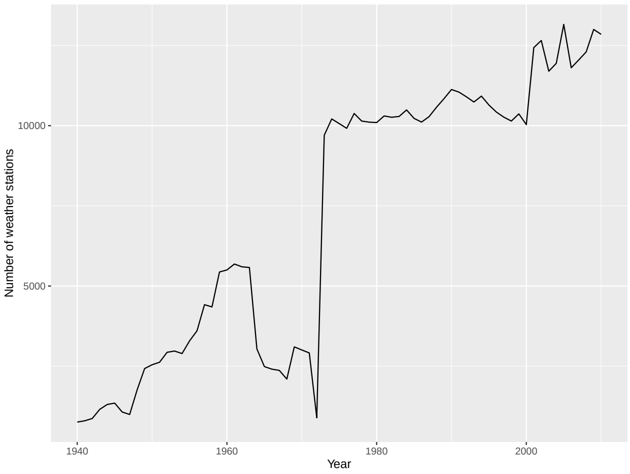

**Figure 1: Total number of weather stations in the ISD data set (1940-2010).**

The initial loading of the data sets, which includes the transformation of the original data format into a more easy to work format, has been separated in two ETL processes, one for the weather measurements (section 4.1) and one for the urbanization and population data (section 4.2). The initially loaded data forms the starting point for the experiments.

## 4.1 Weather measurements ETL

The hourly weather measurements, in the ISD data set, are stored per weather station, per year. The textual format, using row offset to specify specific fields, of the files could be read using a custom developed reader. Since the urbanization data is based on yearly measurements, we also calculated yearly measurements from the hourly measurements. This process, with indications of Spark operations, is illustrated in figure 2A.

In this ETL process we start by converting the custom textual format of the data set to CSV file, resulting in a new file for each station. This file contains one yearly measurement, calculated from the hourly measurements. We keep the identifier of the weather stations, the latitude/longitude coordinates and temperature (min, Q1, median, Q3, max

[2] `https://www.ncdc.noaa.gov/`

[3] `https://www.un.org/development/desa/en/`

[4] `https://github.com/harsha2010/magellan`

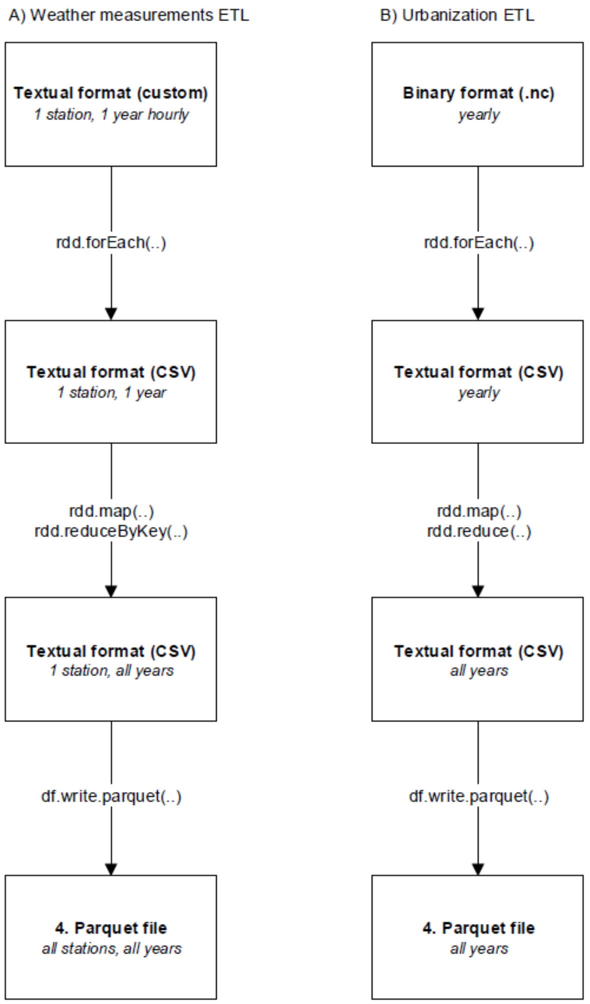

**Figure 2: Schematic overview of the ETL processes.**

and mean). The data set also provides more specific measurements, such as humidity, wind direction and in some cases even rain- and/or snowfall, but for our research the temperature data suffices. These CSV files with yearly measurements are than combined, per station, to files that contain all yearly measurements for each station. Ultimately we combined all the files into a single Parquet files, this file format provides a performance boost if we want to query using Spark SQL, for example from the Spark Shell. We achieved this by reading the CSV files with Spark SQL[5] and then using the build-in functionality to persist a *DataFrame* as a Parquet file.

## 4.2 Urbanization ETL

The land usage data is stored, in one file per year, in the binary format NetCDF to read this data we made use of a reader of Unidata's THREDDS project[6]. The files contain several 2D- and 3D-arrays that represent a grid of world map tiles. Each tile corresponds to a 0.5 by 0.5 degree latitude/longitude tile, using the WGS84 coordinates system[7].

The ETL process starts by converting the binary NetCDF format to CSV files. This had to be done by copying over the files to the Spark nodes, since the reader accepts only file system paths, thus this part of the process could not benefit from the in-memory approach of Spark. After the creation of the CSV file, one per year, these are combined to one single CSV file containing all years. As with the weather measurements the file is converted to a Parquet file. The population data set is already in CSV format and did not require any specific ETL process other than converting the file to a Parquet file.

For each world map tile we used the specified total amount of square meters and and the percentage of urban land coverage to calculate the amount of square kilometers within the world map tile. This is supplemented with the population data and, by combining the two, the population density.

# 5. EXPERIMENTS

Two experiments have been conducted to answer the research questions discussed in 3. The results of the experiments are discussed in this section.

## 5.1 Weather stations in urban areas

To determine the amount of weather stations in urban areas, we first have to determine what the urban areas are in the first place. To achieve this we performed a few additional ETL steps, starting with the files that resulted from the Urbanization ETL process. First we converted the urbanization CSV files to GeoJSON[8] files, one for each year, containing a polygon for each urban area. We did the same for the weather measurements files, resulting in one file per year, containing points for each weather station. Next, for each year the pair of GeoJSON files are loaded using Magellan. Then, they have been joined using Spark SQL and the Magellan *within* expression. This produces a new table with only the the weather stations that are located within an urban area. The number of rows match the number of "urban" weather stations.

[5] `https://github.com/databricks/spark-csv`
[6] `https://github.com/Unidata/thredds`
[7] `http://code7700.com/wgs84.html`
[8] `http://geojson.org/geojson-spec.html`

**Table 1: Weather stations in urban areas (1950-2010)**

| Year | Urban areas (#) | Weather stations (#) | Weather stations in urban areas (#) | Weather stations in urban areas (%) |
|---|---|---|---|---|
| 1955 | 807 | 3174 | 402 | 12.67 |
| 1960 | 876 | 5385 | 550 | 10.21 |
| 1965 | 976 | 2454 | 430 | 17.52 |
| 1970 | 1046 | 2984 | 338 | 11.33 |
| 1975 | 1139 | 9943 | 1224 | 12.31 |
| 1980 | 1198 | 9888 | 1295 | 13.10 |
| 1985 | 1279 | 9833 | 1298 | 13.20 |
| 1990 | 1317 | 10588 | 1387 | 13.10 |
| 1995 | 1358 | 10413 | 1386 | 13.31 |
| 2000 | 1381 | 9513 | 1332 | 14.00 |
| 2005 | 1397 | 12826 | 1674 | 13.05 |
| 2010 | 1397 | 12630 | 1709 | 13.53 |

The results of this experiment are summarized, in steps of five years, in table 1. From this table we can see that on average only one-eight of all the weather stations is located in urban areas. Remember that urban areas are based the 0.5 by 0.5 degree tiles from the land usage data set. These tiles cover, on average, 1670km$^2$ totally. Most of the time, the urban area does not span the whole world map tile. This means that for even the limited amount of urban areas, the weather stations are not necessarily at the center of urban land, but could also be at the borders of the world tile, e.g. relatively "close-by" a city.

Based on this analysis we conclude that these proportion of weather station in urban areas are not an evident sign that the overall picture of global warming calculations is being influenced by weather stations from urban areas, even under the assumption that all of these weather stations are located on places that are subject to UHIs.

## 5.2 Temperature and urbanization correlation

This experiment builds on the information gathered in the previous experiment. We now know which weather stations are located in urban areas and need to compare the temperature measurements of these weather stations with the urbanization information of the corresponding 0.5 by 0.5 world map tile. We used the Spearman rank correlation function in Spark MLlib to determine the strength of the relation between the temperature and each of the urbanization indicators, as presented in section 3. The Spearman method has been used because it doesn't assume the associations between the variables to be linear. As is the case with the increase of urbanization and temperature measurements.

We started, using Spark SQL, by combing all the information from the first experiment in one single table, containing a row for each year a weather station is in a urban area. Since multiple weather stations exist in a single urban area, we calculated the average temperatures per year by combining measurements from all the weather stations within a single urban area world map tile. Then, we supplemented this table with urbanization information, for all the years available: urban land coverage, population and population. It appeared that not all urban areas have a long record of weather measurements, for example, only 51 urban world map tiles have data for the whole 1950-2010 time period. We chose to use 40 as the minimum amount of years a urban area has to have the required data available to have its correlations calculated. This threshold, resulting in 435 urban area tiles, is the result of a trade-off between the length of years, and the amount of urban world map tiles covered.

The overall top 10 highest correlations between temperature and urbanization is listed in table 2. For each indicator the individual top three is indicated in bold. Of the 435 urban areas considered 54 (12.4%) have a "strong" relationship with $r \geq 0.6$ for at least one of the urban indicators. And three (0.7%) areas have a "very strong" relationship ($r \geq 0.8$). Overall, the average correlation is around 0.35 indicating a "weak" relationship. Based on this analysis we conclude that, although a few strong relationships exist, again no sign of influence of urbanization on weather station measurements.

# 6. VISUALIZATION

The data used during this research has been made interactively by means of a interactive world map visualization. It contains all the yearly weather station measurements and urbanization information. Filter can be applied to easily explore the large data set and search for interesting cases within the data. The visualization is accessible online[9] and also contains download links for the GeoJSON data that have been created during the experiments.

# 7. CONCLUSIONS

In this paper we have made documented our findings of studying the influence of urbanization on weather station measurements. The experiments conducted during this study yield no signs that global warming and climate change calculations are heavily influenced by urbanization. Nevertheless, there are weather stations located in urban areas that also have a "strong" relationship between increasing temperature measurements and urbanization indicators. The amounts are practically low and not likely to have any substantial impact on global warming and climate change calculations.

It was challenging to determine the intensity of urbanization based solely on publicly and accessible data. In particular the resolution of the urban land usage, the 0.5 by 0.5 degree resolution seemed to be too course-grained. Since a single tile could contain multiple urban areas, resulting in the situation where it is hard to distinguish between multiple small towns and one larger city.

[9] `https://onnovalkering.github.io/urban-weather`

**Table 2: Top 10 cities with highest temperature-urbanization correlations.**

| City | Urban land with temperature ($r$) | Population with temperature ($r$) | Population density with temperature ($r$) |
|---|---|---|---|
| Singapore, Singapore | 0.859 **(1)** | 0.861 **(1)** | 0.842 **(1)** |
| Valencia, Venezuela | 0.807 **(2)** | 0.806 **(2)** | 0.721 |
| Tianshui, China | 0.787 **(3)** | 0.781 | 0.768 **(3)** |
| Oita, Japan | 0.779 | 0.780 | 0.756 |
| Matsuyama, Japan | 0.320 | 0.804 **(3)** | 0.804 **(2)** |
| Jinhzou, China | 0.707 | 0.704 | 0.741 |
| Hanzhong, China | 0.672 | 0.669 | 0.747 |
| Phoenix, United States | 0.767 | 0.772 | -0.125 |
| Harbin, China | 0.743 | 0.738 | 0.037 |
| Mudanjiang, China | 0.741 | 0.739 | -0.225 |

# 8. REFERENCES

[1] J. Cook, N. Oreskes, P. T. Doran, W. R. Anderegg, B. Verheggen, E. W. Maibach, J. S. Carlton, S. Lewandowsky, A. G. Skuce, S. A. Green, et al. Consensus on consensus: a synthesis of consensus estimates on human-caused global warming. *Environmental Research Letters*, 11(4):048002, 2016.

[2] L. Kleerekoper, M. van Esch, and T. B. Salcedo. How to make a city climate-proof, addressing the urban heat island effect. *Resources, Conservation and Recycling*, 64:30–38, 2012.

[3] K. Klein Goldewijk, A. Beusen, G. Van Drecht, and M. De Vos. The hyde 3.1 spatially explicit database of human-induced global land-use change over the past 12,000 years. *Global Ecology and Biogeography*, 20(1):73–86, 2011.

[4] J. Liu, E. Racah, Q. Koziol, and R. S. Canon. H5spark: Bridging the i/o gap between spark and scientific data formats on hpc systems.

[5] A. M. Rizwan, L. Y. Dennis, and L. Chunho. A review on the generation, determination and mitigation of urban heat island. *Journal of Environmental Sciences*, 20(1):120–128, 2008.

[6] D. Satterthwaite. The implications of population growth and urbanization for climate change. *Environment and Urbanization*, 21(2):545–567, 2009.

[7] A. Smith, N. Lott, and R. Vose. The integrated surface database: Recent developments and partnerships. *Bulletin of the American Meteorological Society*, 92(6):704, 2011.

[8] C. M. Tate, T. F. Cuffney, G. McMahon, E. M. Giddings, J. F. Coles, and H. Zappia. Use of an urban intensity index to assess urban effects on streams in three contrasting environmental settings. In *American Fisheries Society Symposium*, volume 47, pages 291–315. Citeseer, 2005.

[9] United Nations, Department of Economic and Social Affairs, Population Division. World Urbanization Prospects: The 2014 Revision, Highlights (ST/ESA/SER. A/352). *New York, United Nations*, 2014.

[10] J. Zheng and A. Dagnino. An initial study of predictive machine learning analytics on large volumes of historical data for power system applications. In *Big Data (Big Data), 2014 IEEE International Conference on*, pages 952–959. IEEE, 2014.