Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Master Thesis

Improving the Availability of Contextual Data with Machine Learning-Based Interpolation

Ian Maurice Buck

ian.buck@studium.uni-hamburg.de
Study Program Information Systems
Matr.-Nr. 6911467

First Reviewer: Prof. Dr. Janick Edinger
Second Reviewer: Philipp Kisters

Abgabe: 08.2023

A distributed system is one where the failure of some
computer I've never heard of can keep me from getting my work done.
– *Leslie Lamport*

Abstract

Many science-based applications need continuous or gridded input data in order to work properly. This paper investigates how single data-points can be combined to create a continuous data layer, in which missing data points are interpolated. An example for such an application would be the prediction of temperatures coming from single sensors and weather-stations, that can be combined to detect Urban Heat Island (UHI)s. UHIs are weather phenomena that get ampflified among other things by the ongoing densification of urban areas to create more living space, typically accompanied by the removal of green areas that can help with the dissipation of heat. These heat islands, in which the temperature is significantly higher than in surrounding areas, pose a threat to the health of the urban population, especially to the elderly, children and people with existing health conditions. Traditionally, UHIs are detected using Land Surface Temperatures (LST) captured by satellites, that usually have the downside of low spatial and temporal density. This paper proposes an alternative approach by creating a machine-learning model that is able to interpolate missing data between data-points coming from citizen-owned sensor networks that are combined with mobile sensors, which can be attached to rental bikes, buses or e-scooters, to gain temporary insights into otherwise unobserved areas. The model combines data streams of sensor readings with historic data and creates a fine-granular continuous data-layer, in this case for temperature, which allows for an accurate localization of UHIs.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

LST	Land Surface Temperatures
UHI	Urban Heat Island

1 Introduction

In 2023 56% of the human population already lives in urban areas with the number projected to continiously increase to 68% by 2050 [1]. Combined with the ongoing climate change and urban densification, cities are facing many new challenges. With the removal of vegetation in favor of living space and the sealing of surfaces with heat-absorbing materials such as asphalt or concrete for streets and highways [GRGTDW20], rising temperatures lead to new phenomena that pose risks for the urban citizens. A recent phenomenon is the appearance of so called urban heat-islands (UHI). A heat-island is a local occurrence of significantly higher temperatures than surrounding areas that pose a health risk, especially for the elderly, children or citizen with prior health-issues [MBG15].
In order to detect UHIs, Land Surface Temperature (LST) is commonly used. While allowing for a cheap analysis of large areas without the need of ground weather-stations, this approach comes with certain downsides, such as low temporal and spatial resolution and restrictions such as only being able to measure temperatures when no clouds interfere with the microwaces send from the measuring satellite [ZPL15]. This spatial and temporal resolution, of f.e. spatial resolution of 0.01° longitude and 0.01° latitude (equal to roughly 1.11km by 1.11km) and temporal resolution of monthly average surface temperature as offered by LST data provided by the European Space Agency (ESA) Climate Office's data set [GVP], is not enough to effectively analyze the urban microclimate. Another candidate that comes to mind are weather stations. They usually provide hourly, for current values sometimes even 10 min interval readings of temperature, rain and wind, but don't offer the necessary spatial resolution. Lastly, there is the possibility of deploying sensor networks to closely monitor the climate of the city, but this approach can be quite cost intensive for a large amount of sensors over a long time period [CMY+15]. An alternative that is less costly would be to instead rely on citizen-owned sensor networks from the existing Smart Home and Internet of Things (IoT) infrastructure, like Sensor.Community [2] and Netatmo [3], which offer a temporal resolution of 5 min for temperature and wind, hourly for rain, while also having a comparably high spatial resolution. This approach has the desired temporal resolution and has been shown to work well in [MFG+17], but there might be areas, such as industrial zones, where citizens are not able or not allowed to install their personal sensors. In order to also gain insights in such previously unobserved areas, we propose the usage of mobile sensors

[1] https://ourworldindata.org/urbanization#by-2050-more-than-two-thirds-of-the-world-will-live-in-urban-areas
[2] https://deutschland.maps.sensor.community/
[3] https://weathermap.netatmo.com/

that could be installed on buses, bikes or e-scooters to gain temporary snapshots and improve the spatial resolution even further. As research realted activities commonly rely on continuous or gridded data fields, there needs to be a way to convert these single data points from the different sensors into a continuous data-layer.
In this paper, we propose a solution to this problem by training a machine learning regression model, that allows for the interpolation of missing data-points. Based on sensor readings, from the sensor networks and mobile sensors, of commonly collected weather information, such as temperature, humidity, rain, pressure, wind, and possibly other variables such as vegetation indexes [AR20], the model then creates a continuous data-layer that allows for a holistic view of the observed variable, in this case temperature.

To do (??)

1.1 Objective

current situation: abundance of data (data quality unknown), but different data sources at different places with different formats, making it hard to work with different sources at the same time - smart city example -> heat island detection - sensor networks (statinary + mobile) - LST satellite data (lack of spatiotemporal resolution, not suited for microclimate) - vegetation indexes (in geoinformation systems/portals) - etc. - currently these sources are used independently from each other, but how can they be integrated? - hybrid approaches have shown that combining different approaches (smart city stationary + moving sensors) to give better prediction quality than singular approach (reasons: unobserved areas) - statistical models offer not enough flexibility/are to cumbersom to work with (and probabilities are not known) - ML is a good fit to analyse patterns and rules - how can ML be used to integrate the different types of data? - what work needs to be done before using the data in ML (preprocessing, transformation, outlier detection etc.) - one way of preparing data is to interpolate missing data to create continous gridded features (focus of this work) - how ML can we improve the interpolation quality of features? -> turn sparse features into denser versions with interpolation, interpolation based on other features present

In order to validate these things, we need high quality data-sets with many different features, but there are not many available, even tho for certain things (atmosphere etc.) there are OpenData sets available. Goal of this work is to create good data-sets that can be used to train and validate different methods, like ML approaches

1.2 Structure of this work

Research methods:
- literature research as foundation

- heat island detection - smart cities -> sensor networks vs LST - what type of other data is available? geoinformation, vegetatition, for micro-climate: shades of bigger buildings?
- - prototyping
- implement pipeline to pre-process different types of data
- feature extraction
- implement ML model
- deploy ML model

- create/search for fine granular data sets
- add new contextual data to exitsing data sets
- discuss different types of data (gridded vs continous vs data points) and methods for each
- train ML models with different methods (deep learning, random forests etc.) and different features enabled

- cross validation of results
- discuss validation techniques and indicators (RSME, MSE)

The rest of the thesis is structured as follows. Chapter 2 begins with an analysis of related work, where important literature is discussed, that forms the foundation of this research. In chapter 3, the focus lies on describing the service architecture, that shows how a machine learning model can be deployed in different contexts to improve data availability. Which machine learning approaches can be used to interpolate missing data and how they differ from each other is discussed in chapter 4. In Chapter **??**, the different ML approaches are compared and cross validated with each other based on the different model that are trained on the obtained data-sets. Finally, chapter **??** discusses the findings of this thesis and gives an outlook into future work and research directions.

2 Related Work

In the following chapter we lay the foundation for the research conducated in this thesis.

2.1 Interpolation of Missing Data

2.1.1 Regression Analysis in Statistics

- foundation of other research fields, based in statistics/mathematics
  - linear regression (least-squares) - multiple regression models - hierachical regression
  - special cases - piecewise linear regression - inverse prediction - weighted least squares
- logistic regression - poisson regression

2.1.2 Interpolation in Geostatistics

- spatiotemporal (kriging) - time series prediction vs interpolation of missing data - based on GIS - pipeline: fit measured data points to grid, interpolate missing squares

2.1.3 Prediction of Future Data

- Time Series Analysis

2.1.4 Interpolation with Machine Learning Models

- ML regression

2.2 Access to Data-Sets

2.2.1 OpenData Movement

- portals - official: - EU: https://data.europa.eu/data/datasets?query=temperature&locale=en (combinas many governmental and local catalogues) / https://data.europa.eu/data/catalogues?locale=en - USA: https://data.gov/ - UK: https://www.data.gov.uk/ - private: https://www.kaggle.com/
  - strategies: - self procurement (test beds -> Helsinki Testbed (climate research meso sclae), UK Birmingham Testbed (climate + smart city),)

2.3 Applications and Research Areas

- focus on temperature interpolation/climate research

- climate research - high area coverage, low spatio and temporal resolution (5km by 5km squares) - based on LST data (from satelites) -> not the same as air temperature

- micro-climate research - bad/costly area coverage, high spatio and temporal resolution - Urban heat islands - Pollution (fine dust pollution)

- connection to smart cities - integrating many heterogenous data points - detection climatic anomalys - notifiy/warn residents

- important key words

- current status quo

- important authors and current work

-> identify reseach gap - convert single data points to continous/gridded data - improve density of data to gain insights and imrove visibility -> identify areas with low prediction quality

3 System Architecture

The main goal of this paper is to create a service that can convert single data-points into a continuous data-layer by interpolating missing data-points with the help of a machine learning approach. This service can then be used as a building block for other research-related activities/services, in order to abstract the complexity of the data-layer away while providing a good data quality. In the context of temperature interpolation, we deal with three different architectural layers that are shown in Figure 3.1, and expose the service via a publish-subscribe pattern to other services [BJK+19].

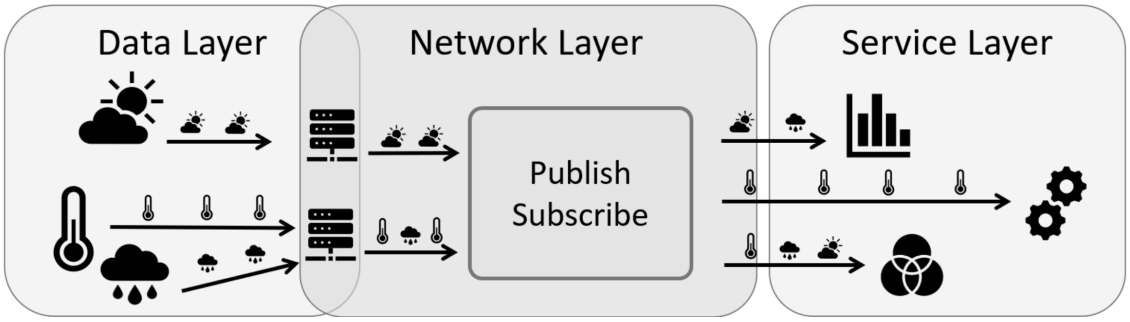

Figure 3.1: In the data layer (left), a wide variety of environmental data is collected with the help of multiple sensors. These are connected to their citizen-owned local base stations, which manage access rights and forward collected data to subscribed services (right) via the decentralized publish-subscribe in the network layer (center).

The *data layer* consists of many different data sources. In the context of temperature sensing and prediction, this could include single (inexpensive) sensors such as the popular BMP280 [1], private weather stations such as sold by Netatmo [2] hidden behind an API [3], public weather station data such as from the Deutscher Wetterdienst (DWD) [4] which offer an API and historic weather data, or other geologically relevant data such as zoning plans which, in the case of the city of Hamburg in Germany, can be accessed via an Open-Data platform provided by the State Office for Geoinformation and Surveying Hamburg [5]. In order to gain detailed insights into urban microclimates, we need fine-grained spatial and temporal data. As managing and maintaining such a large sensor network as a

[1] https://www.bosch-sensortec.com/products/environmental-sensors/pressure-sensors/bmp280/
[2] https://www.netatmo.com/en-gb/weather/weatherstation
[3] https://dev.netatmo.com/apidocumentation/general
[4] https://www.dwd.de/
[5] https://geoportal-hamburg.de/geo-online/

single entity can be quite challenging and cost intensive [CMY+15], we rely primarily on crowdsourced sensor data, in this case climate-related, from citizens, that give access to their personal sensors that they f.e. installed at home. This approach has been shown to work well in the densely populated urban area of Berlin, Germany [MFG+17], with the main challenge being data quality assessment due to faulty data from either broken, wrongly configured or wrongly installed sensors. The different data sources provide data streams which are then ingested by our interpolation service. Main challenges are the uncertainty in networks, such as single sensors or APIs not being available due to network interruptions, and the integration of many heterogeneous data sources that can contain data in different formats, time intervals or units of measurement etc.

The *network layer* is responsible for integrating these different data sources in a consistent and reliable way and making them available for other services. The different sensors present in the data layer might have different vendors and programming interfaces, be located behind (vendor specific) APIs or are unreliably accessible due to unstable networks in edge environments. The network layer can be designed as a peer-to-peer (P2P) network based on the SkipNet approach [HDJ+02], that utilizes the lookup efficiency of distributed hash tables and adds support for value-based range queries based on prefixes and attribute-value pairs. Another challenge in the context of the network layer, especially in the context of this paper, is also the integration of mobile sensors, which might not be constantly connected to a network while moving.

The data layer is then exposed via a publish-subscribe architecture [BJK+19] to the *service layer*, that offers subscriptions to and consumption of data streams and houses services such as our temperature interpolation service. These services can also build upon one another. An example for such a dependency could be a UHI detection service that relies on the temperature interpolation service and offers real-time detection of UHIs, which could trigger notifications/warnings for citizens living in the specific area.

4 Machine Learning Model Design

The data-layer exposes a variety of single data-points for different points in time. As meteorological research and analysis activities are usually in need of gridded or continuous data [SKP+20], in this paper we design a temperature interpolation service that offers continuous temperature data for a given area. This service is part of the *service-layer* and acts as a building block for further temperature related research and analysis, as temperature is an important variable for research in agronomy, meteorology, hydrology, ecology and many other fields of application.
The core of the service is a deployable regression model, that is capable of interpolating missing data points for a target feature in a defined area, based on surrounding data points, that contain various features that are related to the target feature. In the case of temperature interpolation, the target feature is the temperature and the related features could be temperature, rain, humidity, solar radiation or wind, which can be collected using weather stations and specialized sensors. Other features in the context of temperature prediction could also be geological data such as Normalized Difference Vegetation Index (NDVI) or Modified Normalized Difference Water Index (MNDWI), which according to [AR20] have a strong impact on their estimation model.
In order to find the most appropriate model for this application, we want to compare different promising approaches for interpolating data points such as classical geostatistical methods like Empirical Bayesian Kriging (EBK) or EBK-Regression Prediction (EBKRP) [NAEB23] with machine-learning approaches such as random forest regression [AR20]. The goal is to minimize prediction errors and identifying the features that have the biggest impact on prediction quality, like possibly the density of weather stations [NAEB23].
One focus point of this study will also be the underlying uncertainty and dynamic of the data layer. Because the data layer consists of sensors that are connected to a network, there are bound to be times when the network is unreliable and sensors are temporarily not available. This could happen either for bigger areas if a network provider is having an outage in one of its data centers, or more localized, if the Wi-Fi is unreachable for a single sensor. The network is also run by citizens, so each one can decide to turn off their own sensors, exchange them for new ones or add additional sensors to the network. The prediction model of the temperature map service needs to take this into consideration when interpolating the data.

Because we have a highly dynamic underlying data structure that we need to account for, we think that we could improve the prediction results even further by increasing the

Figure 4.1: Some areas are not covered by stationary sensors (left). Whenever mobile sensors collect data in these areas (middle) this knowledge can be used to train regression model which predicts weather conditions for these unmonitored areas (right).

density of data points and adding readings for previously unobserved areas by adding mobile sensors to the data layer. These sensors could be deployed in an urbanized area by being installed on buses, bikes or e-scooters, which also have the advantage that they usually move close to heat absorbing surfaces such as streets or move through parks which can generally help with heat dissipation. The question here would be, what type of sensors are applicable for such local snapshots and which features could be captured. Temperature can be read easily with a small and inexpensive sensor like the BMP280, whereas collecting rain might be difficult for a moving object.

- compare regression based models

5 Preparation of Datasets

- overview of sources for good datasets for temperature and climate related research
goal: - collect mutiple datasets (many features, fine-granualar spaciotemporal) - enhance datasets with additonal information (soil conditions, zoning plans, vegetation health)

Bibliography

[AR20] ALONSO, Lucille ; RENARD, Florent: A new approach for understanding urban microclimate by integrating complementary predictors at different scales in regression and machine learning models. In: *Remote Sensing* 12 (2020), Nr. 15, S. 2434

[BJK+19] BORNHOLDT, Heiko ; JOST, David ; KISTERS, Philipp ; ROTTLEUTHNER, Michel ; BADE, Dirk ; LAMERSDORF, Winfried ; SCHMIDT, Thomas C. ; FISCHER, Mathias: SANE: Smart networks for urban citizen participation. In: *2019 26th International Conference on Telecommunications (ICT)* IEEE, 2019, S. 496–500

[CMY+15] CHAPMAN, Lee ; MULLER, Catherine L. ; YOUNG, Duick T. ; WARREN, Elliott L. ; GRIMMOND, C Sue B. ; CAI, Xiao-Ming ; FERRANTI, Emma J.: The Birmingham urban climate laboratory: an open meteorological test bed and challenges of the smart city. In: *Bulletin of the American Meteorological Society* 96 (2015), Nr. 9, S. 1545–1560

[GRGTDW20] GRÊT-REGAMEY, Adrienne ; GALLEGUILLOS-TORRES, Marcelo ; DISSEGNA, Angela ; WEIBEL, Bettina: How urban densification influences ecosystem services—a comparison between a temperate and a tropical city. In: *Environmental Research Letters* 15 (2020), Nr. 7, S. 075001

[GVP] GHENT, D. ; VEAL, K. ; PERRY, M.: *ESA Land Surface Temperature Climate Change Initiative (LST_cci): Multisensor Infra-Red (IR) Low Earth Orbit (LEO) land surface temperature (LST) time series level 3 supercollated (L3S) global product (1995-2020), version 2.00*. `http://dx.doi.org/10.5285/ef8ce37b6af24469a2a4bdc31d3db27d`

[HDJ+02] HARVEY, Nicholas J. ; DUNAGAN, John ; JONES, Mike ; SAROIU, Stefan ; THEIMER, Marvin ; WOLMAN, Alec: Skipnet: A scalable overlay network with practical locality properties. (2002)

[MBG15] MARTIN, Philippe ; BAUDOUIN, Yves ; GACHON, Philippe: An alternative method to characterize the surface urban heat island. In: *International journal of biometeorology* 59 (2015), S. 849–861

[MFG+17] MEIER, Fred ; FENNER, Daniel ; GRASSMANN, Tom ; OTTO, Marco ; SCHERER, Dieter: Crowdsourcing air temperature from citizen weather stations for urban climate research. In: *Urban Climate* 19 (2017), S. 170–191

[NAEB23] NJOKU, Elijah A. ; AKPAN, Patrick E. ; EFFIONG, Augustine E. ; BABATUNDE, Isaac O.: The effects of station density in geostatistical prediction of air temperatures in Sweden: A comparison of two interpolation techniques. In: *Resources, Environment and Sustainability* 11 (2023), S. 100092

[SKP+20] SEKULIĆ, Aleksandar ; KILIBARDA, Milan ; PROTIĆ, Dragutin ; TADIĆ, Melita P. ; BAJAT, Branislav: Spatio-temporal regression kriging model of mean daily temperature for Croatia. In: *Theoretical and Applied Climatology* 140 (2020), S. 101–114

[ZPL15] ZHANG, Xiaoyu ; PANG, Jing ; LI, Lingling: Estimation of land surface temperature under cloudy skies using combined diurnal solar radiation and surface temperature evolution. In: *Remote Sensing* 7 (2015), Nr. 1, S. 905–921

Eidesstattliche Versicherung

Hiermit versichere ich an Eides statt, dass ich die vorliegende Arbeit im Studiengang XXX selbstständig verfasst und keine anderen als die angegebenen Hilfsmittel – insbesondere keine im Quellenverzeichnis nicht benannten Internet-Quellen – benutzt habe. Alle Stellen, die wörtlich oder sinngemäß aus Veröffentlichungen entnommen wurden, sind als solche kenntlich gemacht. Ich versichere weiterhin, dass ich die Arbeit vorher nicht in einem anderen Prüfungsverfahren eingereicht habe und die eingereichte schriftliche Fassung der auf dem elektronischen Speichermedium entspricht.

Hamburg, den _____ Unterschrift: _____

Veröffentlichung

Ich stimme der Einstellung der Arbeit in die Bibliothek des Fachbereichs Informatik zu.

Hamburg, den _____ Unterschrift: _____