Datasheets for Datasets

# Motivation for Dataset Creation

## **Why was the dataset created?** (e.g., were there specific tasks in mind, or a specific gap that needed to be filled?)

The dataset addresses the need for cross-sectional data for each of Hamburg’s district. The so-called “Stadtteilprofile” (district profiles) are compiled to help actors in politics, the administration, institutions and clubs to make more informed decisions. Since these documents merely collect multiple different statistics for publication, I do not think that they are created with a more specific purpose in mind than to provide accurate information available to the public.

## **What (other) tasks could the dataset be used for?** Are there obvious tasks for which it should *not* be used?

One obvious task would be to quantify the attractiveness of a district based on some of the metrics that are contained in the profiles. This could include business opportunity or helping people find the right neighborhood to look for housing in. One aspect that the data could have a negative impact on is gentrification, as such data might make it even easier for communities to self-select and further drive up rents in certain areas.

## **Has the dataset been used for any tasks already?** If so, where are the results so others can compare (e.g., links to published papers)?

The district profiles have been used by several media outlets to produce spatial graphics that visualize some of the metrics, such as population density or crime rates in Hamburg’s different regions. I was able to find only a handful of publications that cite usage of this data, for example an [analysis of rent and property prices in a newly developed neighborhood](https://books.google.com/books?hl=de&lr=&id=6Y5kAQAAQBAJ&oi=fnd&pg=PR3&ots=ct07AtwPIX&sig=Pln_wq-dWbmpbYjrFY6ZZMoA4oA#v=onepage&q&f=false) and the potential [impact of climate change on different district](https://pure.mpg.de/rest/items/item_2148697/component/file_2465323/content). As far as I can tell, the data has never been published in a form that makes working with it across multiple years easy, thus making it difficult to use the dataset to establish trends.

## **Who funded the creation of the dataset?** If there is an associated grant, provide the grant number.

The dataset is a collection of various sociodemographic features that have been collected by a variety of public institutions and offices in the city and on a federal level. I am very certain that no active data collection was done, in the sense that the district profiles only contain data that have been provided to government agencies by for example citizens registering themselves in a particular area, registering their car, registering a business, etc. The aggregation across administrative units and the publication of the data is paid for by the city and state of Hamburg, which funds the provider [www.statistik-nord.de](http://www.statistik-nord.de/).

## **Any other comments?**

# Dataset Composition

## **What are the instances?** (that is, examples; e.g., documents, images, people, countries) Are there multiple types of instances? (e.g., movies, users, ratings; people, interactions between them; nodes, edges)

The instances are Hamburg’s 104 districts observed in one-year-intervals. The dataset therefore contains almost 1000 instances, if we count each district-year combination as one instance. If we only count the districts as instances, the data contains 10 observations of the district’s features, one for each year.

## **Are relationships between instances made explicit in the data** (e.g., social network links, user/movie ratings, etc.)?

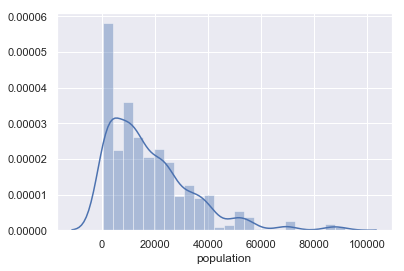
The districts are labelled mutually exclusive and collectively exhaustive, with the second axis being the year of observation.

## **How many instances of each type are there?**

There are 10 observations of 104 districts and some 40 to 50 features per district and observation. The dataset thus consists of about 1,000 rows.

## **What data does each instance consist of?** “Raw” data (e.g., unprocessed text or images)? Features/attributes? Is there a label/target associated with instances? If the instances are related to people, are subpopulations identified (e.g., by age, gender, etc.) and what is their distribution?

After some initial cleaning, the list of features per instance is almost 100 items long (I have described them in my project proposal in detail). Subpopulations are identified where they occur. Here is population histogram in Hamburg’s districts from 2007 to 2017:



## **Is everything included or does the data rely on external resources?** (e.g., websites, tweets, datasets) If external resources, a) are there guarantees that they will exist, and remain constant, over time; b) is there an official archival version. Are there licenses, fees or rights associated with *any* of the data?

The data does not rely on external sources. There is license associated with the data which is held by Statistik Nord which limits the reproduction of the complete dataset.

## **Are there recommended data splits or evaluation measures?** (e.g., training, development, testing; accuracy/AUC)

-

## **What experiments were initially run on this dataset?**

## Have a summary of those results and, if available, provide the link to a paper with more information here.

-

## **Any other comments?**

-

# Data Collection Process

## **How was the data collected?** (e.g., hardware apparatus/sensor, manual human curation, software pro- gram, software interface/API; how were these con- structs/measures/methods validated?)

The data stems from local administrative units which sometimes coincide with districts, sometimes with wards. By registering oneself as being a resident of a certain administrative unit, any citizen also consents to their data being used to prepare reports containing non-individual reports containing the socio-demographic features of that particular administrative unit. This consent only extends to the publication of general and non-identifiable data.

## **Who was involved in the data collection process?** (e.g., students, crowdworkers) How were they compensated? (e.g., how much were crowdworkers paid?)

Since demographic data on a district level is aggregated from the registers of administrative units, the individuals involved on the lowest levels are those working in the local government branches. They enter the data that verified residents of a particular unit provide, such as current address, date of birth, employment status, marital status, etc. The provider of the data, Statistik Nord, only receives aggregate data from the administrative units and prepares reports in which this data is visualized. They are thus not involved in the collection of the data, but can decide which aspects to publish and which to withhold.

## **Over what time-frame was the data collected?** Does the collection time-frame match the creation time-frame?

The data seems to have been collected over one year intervals. It therefore contains some features that have accumulated over the observation period and enter the record at an artificially chosen date, thus not really coinciding with their first observation.

## **How was the data associated with each instance acquired?** Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part of speech tags; model-based guesses for age or language)? If the latter two, were they validated/verified and if so how?

I think the data was acquired through a combination of the above factors and has likely been verified in comparison with other government data.

## **Does the dataset contain all possible instances?** Or is it, for instance, a sample (not necessarily random) from a larger set of instances?

It contains some information for every district-year combination, but the same features for all district-year combinations, leading to some random values. The mean number of instances associated with every feature is approx. 750 out of 1091 possible instances.

## **If the dataset is a sample, then what is the population?**

## What was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)? Is the sample representative of the larger set (e.g., geographic coverage)? If not, why not (e.g., to cover a more diverse range of in- stances)? How does this affect possible uses?

The instances do not encompass every individual living in every district, but all individuals that have turned up in public records over the year of observation. The data can therefore be interpreted as a sample population for the entire district.

## **Is there information missing from the dataset and why?**

## (this does not include intentionally dropped instances; it might include, e.g., redacted text, withheld documents) Is this data missing because it was unavailable?

Some features have not been observed in some years as opposed to others, thus leading to missing values in some columns that have only been observed for a different year.

## **Are there any known errors, sources of noise, or redundancies in the data?**

There are some instances that do not represent districts, but aggregates on a state or ward level.

## **Any other comments?**

-

# Data Preprocessing

## **What preprocessing/cleaning was done?** (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values, etc.)

None

## **Was the “raw” data saved in addition to the preprocessed/cleaned data?** (e.g., to support unanticipated future uses)

The raw data might still be available from the sources the profiles have been compiled from and is often available in a single-feature form

## **Is the preprocessing software available?**

-

## **Does this dataset collection/processing procedure achieve the motivation for creating the dataset stated in the first section of this datasheet?**

-

## **Any other comments?**

# Dataset Distribution

## **How is the dataset distributed?** (e.g., website, API, etc.; does the data have a DOI; is it archived redundantly?)

I was not able to find a DOI for the data, but I did find an ISSN for the latest publication of the district profiles: ISSN 1863-9518

## **When will the dataset be released/first distributed?** (Is there a canonical paper/reference for this dataset?)

ISSN 1863-9518 for the 2017 dataset.

## **What license (if any) is it distributed under?** Are there any copyrights on the data?

*„ Auszugsweise Vervielfältigung und Verbreitung mit Quellenangabe gestattet. Sofern in den Produkten auf das Vorhandensein von Copyrightrechten Dritter hingewiesen wird, sind die in deren Produkten ausgewiesenen Copyrightbestimmungen zu wahren. Alle übrigen Rechte bleiben vorbehalten.“*

Partial distribution or publication allowed if source if provided, except for parts that are copyrighted by other entities. All other rights reserved.

## **Are there any fees or access/export restrictions? Any other comments?**

-

## **Any other comments?**

-

# Dataset Maintenance

## **Who is supporting/hosting/maintaining the dataset?**

## How does one contact the owner/curator/manager of the dataset (e.g. email address, or other contact info)?

The entity responsible for hosting the data is Statistik Nord. Their contact information is available on their website:

<https://www.statistik-nord.de/service-ueber-uns/informationsservice/>

Their contact email address is [info@statistik-nord.de](mailto:info@statistik-nord.de)

## **Will the dataset be updated?** How often and by whom? How will updates/revisions be documented and communicated (e.g., mailing list, GitHub)? Is there an erratum?

The dataset as a whole is not being updated, but new profile data is published on a yearly basis.

## **If the dataset becomes obsolete how will this be communicated?**

-

## **Is there a repository to link to any/all papers/systems that use this dataset?**

-

## **If others want to extend/augment/build on this dataset, is there a mechanism for them to do so?** If so, is there a process for tracking/assessing the quality of those contributions. What is the process for communicating/distributing these contributions to users?

-

## **Any other comments?**

-

# Legal and Ethical Considerations

**If the dataset relates to people (e.g., their attributes) or was generated by people, were they informed about the data collection?** (e.g., datasets that collect writing, photos, interactions, transactions, etc.)

Since the data stems from the public administration sector and contains mostly demographic data, people must have been aware about the data being collected. By registering oneself with the local district administration, which has to be done within the first month of moving into a new German administrative unit, the citizen also consents to being included in demographic reports that the administrative unit can publish at any time.

**If it relates to other ethically protected subjects, have appropriate obligations been met?** (e.g., medical data might include information collected from animals)

-

**If it relates to people, were there any ethical review applications/reviews/approvals?** (e.g. Institutional Review Board applications)

-

**If it relates to people, were they told what the dataset would be used for and did they consent? What community norms exist for data collected from human communications?** If consent was obtained, how? Were the people provided with any mechanism to revoke their consent in the future or for certain uses?

-

**If it relates to people, could this dataset expose people to harm or legal action?** (e.g., financial social or otherwise) What was done to mitigate or reduce the potential for harm?

This dataset could not expose people to harm or legal action, as it does not contain any information related to specific individuals.

**If it relates to people, does it unfairly advantage or dis- advantage a particular social group?** In what ways? How was this mitigated?

-

**If it relates to people, were they provided with privacy guarantees?** If so, what guarantees and how are these ensured?

-

**Does the dataset comply with the EU General Data Protection Regulation (GDPR)?** Does it comply with any other standards, such as the US Equal Employment Opportunity Act?

Since coming into effect, any data that is not stored compliant with the GDPR had to be deleted. Any data found in the district profiles has been collected by those administrative units where residents have to register themselves at. The information gathered in this way has to be GDPR compliant, as the regional authorities would not be able to collect these data otherwise.

**Does the dataset contain information that might be considered sensitive or confidential?** (e.g., personally identifying information)

-

## **Does the dataset contain information that might be considered inappropriate or offensive?**

-

## **Any other comments?**