

Datasheet for Counting the Costs City Data dataset.*

Rui Hu

2001-12-03

AidData has partnered with the Municipal Finance and Urban Economy branch of UN-Habitat, the UN’s agency for human settlement, to conduct a pilot study of the new methodology in cities across five countries—Bolivia, Colombia, India, Malaysia, and Sweden—that are representative of developed and developing contexts. In this report, we outline our reproducible methodology to systematically quantify the costs for cities of achieving SDG 11 across five thematic areas—housing, transportation, waste management, public spaces, and governance—and provide country briefs with our methodology’s results for each country. This dataset contains the data used in our Counting the Costs methodology. In the next phase of our work, we hope to refine and scale the methodology to include other thematic areas and develop decision support tools that help cities maximize their resource allocation.

Motivation

1. *For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.*
 - This dataset contains city data for five countries—Bolivia, Colombia, India, Malaysia, and Sweden—used in our Counting the Costs report to quantify the city-level costs for urban areas to achieve SDG11.
2. *Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, company, institution, organization)?*
 - The dataset was created by the AidData research lab, which operates as part of the College of William & Mary.
3. *Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.*

*Code and data are available at: [LINK](#).

- The dataset was funded by AidData’s institutional resources. The names are Prakash, M., Kamiya, M., Cheng, M., & S. Schaedel.
4. *Any other comments?*
- NA

Composition

1. *What do the instances that comprise the dataset represent (for example, documents, photos, people, countries)? Are there multiple types of instances (for example, movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.*
 - Each instance in the dataset represents a city in a developing country. Data points include economic indicators, measures of conflict, population, and infrastructure characteristics.
2. *How many instances are there in total (of each type, if appropriate)?*
 - The dataset includes 151 cities.
3. *Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (for example, geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (for example, to cover a more diverse range of instances, because instances were withheld or unavailable).*
 - The dataset is a sample focused on cities from developing countries, with coverage determined by data availability. It is not exhaustive but aims to capture a diverse range of urban environments.
4. *What data does each instance consist of? “Raw” data (for example, unprocessed text or images) or features? In either case, please provide a description.*
 - Each instance contains city-level metrics such as households, income, rent, and population density. The data has been processed to provide meaningful indicators.
5. *Is there a label or target associated with each instance? If so, please provide a description.*
 - NA
6. *Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (for example, because it was unavailable). This does not include intentionally removed information, but might include, for example, redacted text.*
 - Some cities have incomplete data due to limitations in available sources, such as Current Length of Roads or Paved roads (% total).

7. *Are relationships between individual instances made explicit (for example, users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.*
 - No explicit relationships are defined between cities; however, spatial patterns or correlations could be inferred using geolocation data.
8. *Are there recommended data splits (for example, training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.*
 - NA
9. *Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.*
 - NA
10. *Is the dataset self-contained, or does it link to or otherwise rely on external resources (for example, websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (that is, including the external resources as they existed at the time the dataset was created); c) are there any restrictions (for example, licenses, fees) associated with any of the external resources that might apply to a dataset consumer? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.*
 - Yes.
11. *Does the dataset contain data that might be considered confidential (for example, data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.*
 - No. The dataset consists of publicly available information.
12. *Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.*
 - No. The data is neutral and focused on quantitative analysis.
13. *Does the dataset identify any sub-populations (for example, by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.*
 - No.
14. *Is it possible to identify individuals (that is, one or more natural persons), either directly or indirectly (that is, in combination with other data) from the dataset? If so, please describe how.*

- No, the dataset does not include individual-level data
15. *Does the dataset contain data that might be considered sensitive in any way (for example, data that reveals race or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.*
- No, the data is aggregated at the city level and does not include sensitive individual information.
16. *Any other comments?*
- NA

Collection process

1. *How was the data associated with each instance acquired? Was the data directly observable (for example, raw text, movie ratings), reported by subjects (for example, survey responses), or indirectly inferred/derived from other data (for example, part-of-speech tags, model-based guesses for age or language)? If the data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.*
 - Data was gathered from multiple public sources, including national economic statistics.
2. *What mechanisms or procedures were used to collect the data (for example, hardware apparatuses or sensors, manual human curation, software programs, software APIs)? How were these mechanisms or procedures validated?*
 - A combination of manual data curation and automated extraction techniques from trusted repositories and datasets.
3. *If the dataset is a sample from a larger set, what was the sampling strategy (for example, deterministic, probabilistic with specific sampling probabilities)?*
 - The dataset includes cities with available data on conflict and economic indicators, focusing on developing countries.
4. *Who was involved in the data collection process (for example, students, crowdworkers, contractors) and how were they compensated (for example, how much were crowdworkers paid)?*
 - AidData researchers and analysts conducted the data collection. Compensation details are not specified.

5. *Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (for example, recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.*
 - Not explicitly mentioned in the dataset documentation.
6. *Were any ethical review processes conducted (for example, by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.*
 - Not explicitly mentioned in the dataset documentation.
7. *Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (for example, websites)?*
 - The data was obtained from third-party sources from AIDDATA.
8. *Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.*
 - Not applicable, as no individual-level data is included.
9. *Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.*
 - TBD
10. *If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).*
 - Not applicable.
11. *Has an analysis of the potential impact of the dataset and its use on data subjects (for example, a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.*
 - No specific analysis of this nature is mentioned.
12. *Any other comments?*
 - NA

Preprocessing/cleaning/labeling

1. *Was any preprocessing/cleaning/labeling of the data done (for example, discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remaining questions in this section.*
 - Yes, we select several predictors of original dataset.
2. *Was the “raw” data saved in addition to the preprocessed/cleaned/labeled data (for example, to support unanticipated future uses)? If so, please provide a link or other access point to the “raw” data.*
 - Yes. https://github.com/XYPKQ896/housing_price/blob/main/data/01-raw_data/original_data.csv
3. *Is the software that was used to preprocess/clean/label the data available? If so, please provide a link or other access point.*
 - Rstudio.<https://posit.co/>.
4. *Any other comments?*
 - NA

Uses

1. *Has the dataset been used for any tasks already? If so, please provide a description.*
 - Yes, it has been used for analyzing drivers of Urban Housing Prices.
2. *Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.*
 - NA.
3. *What (other) tasks could the dataset be used for?*
 - Quantify the city-level costs for urban areas to achieve SDG11
4. *Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a dataset consumer might need to know to avoid uses that could result in unfair treatment of individuals or groups (for example, stereotyping, quality of service issues) or other risks or harms (for example, legal risks, financial harms)? If so, please provide a description. Is there anything a dataset consumer could do to mitigate these risks or harms?*
 - Data gaps and inconsistencies may limit certain analyses.
5. *Are there tasks for which the dataset should not be used? If so, please provide a description.*

- NA
6. *Any other comments?*
- NA

Distribution

1. *Will the dataset be distributed to third parties outside of the entity (for example, company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.*
 - Yes, the dataset is openly accessible via AidData's website
2. *How will the dataset be distributed (for example, tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?*
 - The dataset is available as a downloadable file.
3. *When will the dataset be distributed?*
 - It is already publicly available.
4. *Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/ or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.*
 - Yes, the dataset is subject to AidData's terms of use.
5. *Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.*
 - None mentioned.
6. *Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.*
 - None mentioned.
7. *Any other comments?*
 - NA

Maintenance

1. *Who will be supporting/hosting/maintaining the dataset?*

- AidData AidData (2024)
2. *How can the owner/curator/manager of the dataset be contacted (for example, email address)?*
 - Through the AidData website.
 3. *Is there an erratum? If so, please provide a link or other access point.*
 - Not mentioned.
 4. *Will the dataset be updated (for example, to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to dataset consumers (for example, mailing list, GitHub)?*
 - Updates are not explicitly mentioned.
 5. *If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (for example, were the individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.*
 - NA
 6. *Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to dataset consumers.*
 - Not specified.
 7. *If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to dataset consumers? If so, please provide a description.*
 - Swedish International Development Agency (SIDA).
 8. *Any other comments?*
 - NA

References

AidData. 2024. “Counting the Costs: City-Level Data.” <https://www.aiddata.org/data/counting-the-costs-city-data>.