4th October 2021

MIGRATION DATASET

datasheet

1. motivation  
   1. For what purpose was the dataset created?   
        
      The Migration Dataset was created to provide population change information of three regions of interest- Canterbury, Wellington and Auckland, to study the problem statement: How does the severity of earthquakes in Canterbury, New Zealand affect migration of residents and their mental health. The dataset was created for the specific task to be combined with two other datasets namely Mental Health Consultation Dataset and Earthquake Severity for the purpose of the study.
   2. Who created this dataset?  
        
      The original dataset is sourced from the Stats NZ Tatauranga Aotearoa's (Stats NZ) online database site (<http://infoshare.stats.govt.nz/>) on the purpose of providing public access to statistics of various measures and topics of the New Zealand region. The NZ Stats site is a central location for accessing New Zealand data and statistics.  
        
      Using the dataset provided publicly by the Ministry of Health, the DATAMAC team consisting of Chenthi Heer, Maple (Jeong Im) Lee, Zhenyuan He and Ada (Rabbiatul Adawiyah) Mohd Izhair of the DATAMAC group have cleaned up the original dataset for specific use of the DATA201 project for the University of Canterbury.
   3. Who funded the creation of this dataset?  
        
      This dataset did not require any direct funding to the DATAMAC team as all data extracted from Stats NZ was data open for public access for free without any costs.

2. Composition

1. What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.  
     
   The instances are of population numbers according to age group and gender in the Canterbury, Wellington and Auckland region.
2. 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 (e.g., 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 (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).  
     
   The dataset contains all possible instances of population numbers within the dataset period (January 2005- December 2020) in the three regions of interest.
3. Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.  
     
   This dataset contains no missing or NULL values from individual instances of earthquake data.

3. collection process

1. 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 data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.  
     
   The data was directly observable as it consists of the population count for the three regions.
2. What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? How were these mechanisms or procedures validated?  
     
   Unspecified.
3. Who was involved in the data collection process?  
     
   The data collection process involved government officials working under Statistics NZ.
4. Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.  
     
   This specific dataset is over the timeframe of 15 years, from 2005 until the 2020 financial year.
5. Were any ethical review processes conducted (e.g., 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.  
     
   Ethical review processes were not specified in the original dataset, however is highly trusted due to it being a governmental institution.
6. Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?  
     
   Via the Statistics NZ site, link provided in 1(B).
7. Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted?   
     
   Potential impacts the dataset could bring towards data subjects have been discussed however due to the dataset not including any private information on individual patients, we believe this dataset does not possess any threat to the data subjects in question. The dataset only contains the count of population in the regions of interest (Christchurch, Wellington, and Auckland).

4. USES

1. Has the dataset been used for any tasks already? If so, please provide a description.  
     
   Aligned with the purpose of this dataset being created as stated in section A, this dataset has been used to investigate the relationship of population changes with a rough measure of earthquake severity.   
     
   The dataset is included in an R package that can be installed by public through the "remotes" package and the github repository "chenthih".
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.  
     
   Yes, on github. [*https://github.com/adaizh/datamac*](https://github.com/adaizh/datamac)

5. Distribution

1. Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.   
     
   No, the dataset will not be distributed to third parties outside the entity.
2. How will the dataset will be distributed (e.g., tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?  
   The dataset does not have a digital object identifier, but is accessible via github under the repository "adaizh/datamac".
3. When will the dataset be distributed?  
     
   The dataset is available on the github repository as of this writing.