exploratory

William Gerecke

1/27/2022

Abstract

This is my abstract.[1]

Contents

[1] footnote

1 Intro

abbreviations: - greenhouse gas (GHG)

TODO: - data cleaning: don't hard code row counts - check for past/present tense

2 Data

2.1 Software

This project uses the statistical programming language R (R Core Team 2021) to generate graphics and process data. The data is managed using `tidyverse` and processing is done using `dplyr` verbs (Wickham et al. 2019, 2021). For geometry-specific processing, the `sf` package is used (Pebesma 2018). The data is downloaded from the Open Data Toronto Portal using the `opendatatoronto` package (Gelfand 2020). The `tidygeocoder` package is used to geocode addresses present in the downloaded data set (Cambon et al. 2021).

2.2 Source

Regarding this project there are two data sets of interest, both of which come from the Open Data Toronto Portal. The first is the Annual Energy Consumption data set which contains columns for the energy consumption of individual buildings in Toronto that are required by Ontario Regulation 397/11, the Green Energy Act (2009), to report their GHG emissions. Specifically, this data set contains annual `xlsx` spreadsheets from 2011-2018 that have columns with the building: name; address; floor area; mega liters of water or sewage treated; amount of energy purchased in the form of electricity and natural gas; and GHG emissions. The distribution by year of these features can be seen in Figure 1. The features are plotted on a logarithmic scale because these features vary most significantly in their order of magnitude.

The second data set is called City Wards and it contains a shape file (`.shp` extension) that describes the 25 municipal wards of Toronto. The data contains the name of each ward as well as its geospatial information that will be used to group building data by ward.

In order to identify which ward each building belongs to, the building coordinates are required (latitude and longitude), which are obtained by using the `tidygeocoder` package. The `tidygeocoder` package is free and open source, but geocoding API requests are limited to one request per second.

2.3 Cleaning

The Annual Energy Consumption dataset is provided in the form of several `xlsx` spreadsheets with a preamble at the top and grouped cells that describe column labels. The top rows were truncated and columns were manually assigned the correct labels. Additionally, there was one file containing spreadsheets for the years 2011-2014 which had to be separated. There were several paired columns where one contained a number and the other contained a unit. These were appropriately converted to the same unit using standard unit conversions. Finally, some postal codes contained a space in the middle (e.g. `M4Y 0A9`) and some did not. For the purpose of geocoding, these had to be converted to the prior format.

In order to identify which wards each building belongs to, they were geocoded by address, postal code, city, and country. There were 2404 unique addresses, 738 of which could not be geocoded that could not be geocoded for various reasons, such as `'Various addresses'` in the address column. In order to reduce the strain on the Nomination API, only unique addresses were geocoded which reduced the number of necessary calls to the API by 6765. The yearly reports and a summary of the geocoding process can be seen in Table 1.

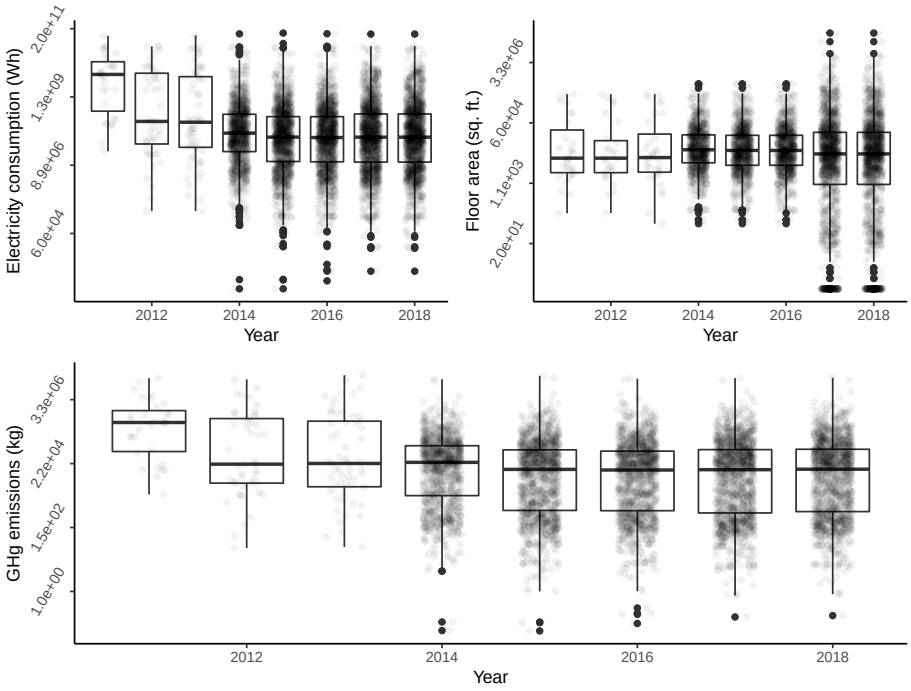

Figure 1: Reports for GHG emissions, electricity consumption, and floor area by year on a logarithmic scale.

Table 1: Number of successful and unsuccessful gecoded addresses for each year.

Geocode status	2011	2012	2013	2014	2015	2016	2017	2018	Total
failure	19	125	46	52	93	95	154	154	**738**
success	492	858	986	1,021	1,210	1,208	1,328	1,328	**8,431**
Total	511	983	1,032	1,073	1,303	1,303	1,482	1,482	**9,169**

Each building in the dataset was assigned a ward based on which ward's boundaries the geocoded point for that building intersected with. At the end of the cleaning process, excess columns were dropped, rows with invalid data were dropped, and the dataset was saved to the `inputs/data/cleaned` directory. The number of buildings that submitted reports in each ward is shown in Figure 2.

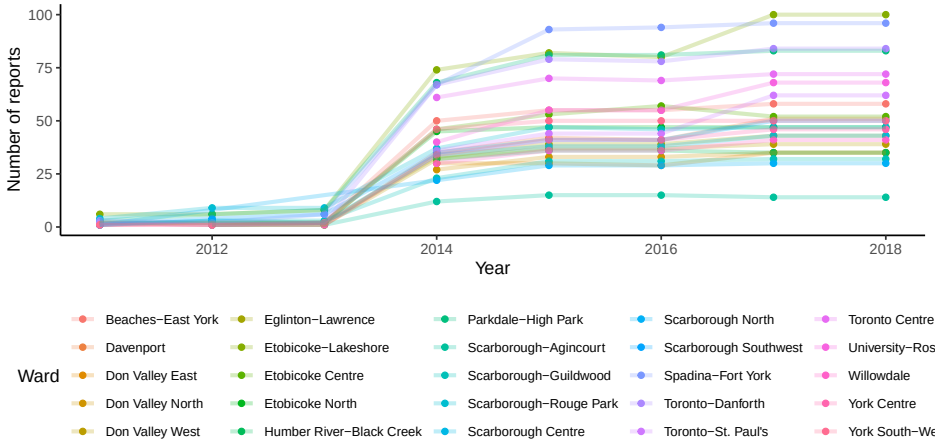

Figure 2: Number of reports each year for each ward.

3 Results

4 Discussion

- geocoding by postal code and not address
  - might assign to wrong ward
- dropped wrong values
- some buildings aren't required to report
  - some buildings chose to report and weren't required to
    - possibly skew results because people wouldn't want to report if they were bad

Cambon, Jesse, Diego Hernangómez, Christopher Belanger, and Daniel Possenriede. 2021. "Tidygeocoder: An r Package for Geocoding." *Journal of Open Source Software* 6 (65): 3544. https://doi.org/10.21105/joss.03544.

Gelfand, Sharla. 2020. *Opendatatoronto: Access the City of Toronto Open Data Portal*. https://CRAN.R-project.org/package=opendatatoronto.

Pebesma, Edzer. 2018. "Simple Features for R: Standardized Support for Spatial Vector Data." *The R Journal* 10 (1): 439–46. https://doi.org/10.32614/RJ-2018-009.

R Core Team. 2021. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, Romain François, Lionel Henry, and Kirill Müller. 2021. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.