Final Report

Abstract

This report presents our efforts supporting Public Health - Seattle and King County (PHSKC) in developing an advanced model for generating parcel-level population estimates in King County. PHSKC serves a diverse population of nearly 2.53 million people across various socioeconomic and cultural backgrounds. The department has noted refining population estimates across communities and zip codes to be of great interest, as these metrics allow for effective resource allocation, disease surveillance, and healthcare planning.

Existing methodologies employed by PHSKC rely on bedroom counts from King County Department of Assessments data. These models result in significant overestimates of population levels when compared to Census tract-level data. Recognizing the limitations of existing models, this report introduces a novel approach utilizing American Community Survey (ACS) data.

Our methodology encompassed a multi-stage process, beginning with the division of ACS data into training and testing/validation sets. We explored two classes of predictive models, multiple linear regression and Poisson regression, to estimate the mean number of people living in a household and the expected number of people per bedroom, respectively. The effectiveness of these models was assessed through cross-validation error comparison. Subsequently, the performance of the model with the lowest cross-validation error was formally assessed on the validation set and that model was applied to parcel-level data obtained from the King County Assessor’s office.

Results showed promise in bridging the gap between ACS Census data and parcel-level estimates. The final phase of our study involved a comparative analysis of our model's predictions against decennial Census data, aiming to validate its accuracy and reliability. This work is expected to contribute to PHSKC's mission by providing more precise population estimates, thereby enhancing public health initiatives and services across King County.

Background

We worked with Public Health - Seattle and King County, which works to protect and improve the health and well-being of all people in King County as measured by increasing the number of healthy years that people live and eliminating health disparities.

PHSKC consists of 1,200 employees serving over 40 different sites and operating with a biennial budget of $1 billion. With service to almost 2.53 million residents, PHSKC is one of the largest public health departments in the United States. The population of King County is diverse, with over 100 languages spoken in the region, and the public health arena reflects this, with over 7,000 medical professionals serving across 19 acute care hospitals and beyond. PHSKC provides countless essential services in support of these populations and medical efforts.[1]

Problem Introduction

For a variety of applications such as resource allocation, disease surveillance and healthcare planning, PHSKC required small area population estimates. Given a small geographical area, the department would like a precise estimate of the number of people living in that area.

Unfortunately, for many applications, the desired data was not available. In most cases, the best publicly available data is U.S. Census Bureau decennial data. However, this data is aggregated geographically and injected with noise before it is released to the public to protect individual privacy.[2] Data is aggregated to several levels, the smallest of which is the Census block. Census blocks are then aggregated to block groups, and then to Census tracts. The Census Bureau advises against the use of data published on the smallest aggregate regions due to the large proportion of injected noise. However, Census tracts typically contained between 1,200 and 8,000 people[3] and thus are far too large for PHSKC's desired use.

In order to make predictions and display results on regions familiar to the public, such as neighborhoods or ZIP codes, PHSKC desired parcel-level estimates of population which could subsequently be aggregated to the desired regions. A parcel is defined as an identifiable unit of land that is treated as separate for valuation or zoning purposes[4]. In our setting, a parcel is a plot of land that contains either a single-family home or a group of apartment or condominium units.

Discussion of Related Work

PHSKC previously attempted to estimate parcel-level populations using data on the number of bedrooms per parcel obtained from the King County Department of Assessments. This dataset provided information on the number of bedrooms for apartments, condominiums, and residential units throughout King County. However, the modeling approach attempted by PHSKC presented two problems:

- Firstly, when aggregating parcel-level population predictions to the Census tract level, the resulting in notably higher estimates to that of the Census Bureau's official estimates, which we consider to be ground truth.
- Additionally, the resulting modeling framework only permitted parcel-level population estimates for the present year. An ideal model should incorporate past historical data and possess the capability to project population figures into both the past and future.

Primary and Secondary Objectives

Our primary objective was to develop a model which accurately predicts the population of regions of interest (parcels) for a single year. Our aim was to build a regression model using known population data, then apply the fitted model to predict the population of parcels for which predictor variables are available but population data is not.

Our secondary objective was to extend our model to enable multi-year population predictions. Due to time constraints, we were not able to produce this multi-year model but instead provide our sponsor with a modeling framework that could be updated to incorporate new data each year.

Methods

ACS Data Acquisition and Cleaning

The American Community Survey was our primary data source for building our predictive models. We utilized the `tidycensus` package in `R` to pull 2021 ACS data from the US Census Bureau's data API.

Because this data was obtained via a complex two-stage sampling design[16], we needed to properly account for survey weights. To this end, we constructed `svydesign` objects in `R` using the `survey` package, which combine a data frame with the survey design information needed to analyze it. Initial data cleaning involved re-coding certain ACS variables to match what was available in the Assessor's data. In particular, the ACS variable representing the number units in a structure (`BLD`) was used to construct a variable (`unit_type`) indicating whether a sampled household is a standalone single-family home (`House`), a condominium (`Condo`), or an apartment (`Apartment`).

Further, because survey data was subject to non-response and conditional response (i.e. questions which were only answered if conditions on previous questions were met), we constructed indicator variables to indicate whether a variable should be considered

for a given observation. In particular, because the ACS variable indicating the tax assessed value of a unit (`TAXAMT`) was not defined for all `unit_types`, we constructed an indicator variable which took value 1 for all observations with defined `TAXAMT` and 0 for all observations with undefined `TAXAMT` (`include_tax`).

Additionally, in order to perform Poisson regression modeling (see Methods - Model Fitting) of bedroom occupancy rate, defined as number of people per bedroom, we defined a new variable `BDSP_nostudio` which took the same value as `BDSP` when that value wasn't 0 and took the value 1 if `BDSP` was 0. This was necessary to avoid infinite rates and our opinion is that studio units and one-bedroom units are not meaningfully different.

Another ACS variable of note is the Public Use Microdata Area (`PUMA`) in which a parcel lies. PUMAs are defined as non-overlapping statistical geographical areas which partition each geographical entity into regions containing no fewer than 100,000 people[9]. There are 16 PUMAs within King County, numbered 11601-11616.

We produced descriptive statistics for relevant covariates in the ACS data (see *Tables* 1-4). Count-type descriptive statistics were produced both prior to and after incorporating survey weights in order to provide a clear picture of our sample size as well as the population which this sample was meant to represent. In order to avoid misleading results, all other descriptive statistics were provided only after incorporating survey weights. It should be noted that descriptive statistics displayed were for only non-group housing settings, as weights for group housing were not included in the ACS data. Further, for any variables which were irrelevant for certain household types, the irrelevant household type was excluded when producing the descriptive statistics (e.g. if tax amount was not relevant for an apartment household, that apartment was excluded from this calculation).

Finally, in order to report final model performance, we sampled 20% of our data, stratified by number of people per household (`NP`), to be held out as a validation set.

King County Assessor's Data Acquisition and Cleaning

We sourced data from the King County Assessor's Office to serve as the input data for the fitted model to estimate the number of people living on each parcel in King County. We sourced this data in several zipped .csv files directly from the King County Assessor's office website[5]. In particular, we utilized the following files: "Apartment

Complex (.ZIP)", "Condo Complex and Units (.ZIP)", "Residential Building (.ZIP)", "Real Property Account (.ZIP)", and "Unit Breakdown (.ZIP)".

The Apartment Complex file contains apartment complex-level characteristics and the Unit Breakdown file contains unit-level characteristics for every apartment complex appearing in the Apartment Complex file. The Condo Complex and Units and Residential Building file each contain unit-level characteristics for unique condominium and single-family homes, respectively.

We used the `tidyverse` package in `R` to import the relevant .csv files. We then merged apartment complex-level data found in the Apartments Complex file with apartment unit-level data found in the Unit Breakdown file according to the unique identifiers "Major" and "Minor". This merged dataset was then combined with the Condo Complex and Units and Residential Building datasets to create a single dataset containing every housing unit in the county. This dataset was then merged with the Real Property Account data in order to obtain tax assessed value for every housing unit in the county.

In order to assign each residential unit to the correct PUMA, we first obtained geographical coordinates for each unit via the `tidygeocoder` package in `R`, then used PUMA shapefiles downloaded from IPUMS USA[18] to determine in which PUMA each address lies. We then added the column `PUMA` to our dataset.

We produced descriptive statistics from this resulting dataset. Counts of the number of the three primary housing units (house, condo, and apartments) are described in *Table 5*. Counts of group housing were not considered as there was no corresponding variable resulting from the ACS data.

Additional summary statistics of beds across county housing units can be found in *Table 6*.

Model Fitting

We fitted several types of models using various subsets of our predictor set (described in "ACS Data Acquisition and Cleaning"). To determine model performance, we utilized a 5-fold cross-validation scheme. That is, we split our training data into five roughly equally sized subsets, fitted each model 5 times to each possible subset of ⅘ of the data, then used each model to predict outcomes for the remaining ⅕ of the data. For each fold, a new `survey` object was fit.

In order to compare models, the root mean squared error was calculated for each model. This error was calculated by first calculating the weighted root mean squared error (RMSE) for each fold (weighted according to the survey weights of the hold-out set), then averaging across all 5 folds. That is, for a given model error was calculated as

$$RMSE_w = \frac{1}{5}\sum_{j=1}^{5}\left(\sqrt{\frac{\sum_{i\in I}^{n} w_i*(y_i-\hat{y}_i)^2}{\sum_{i=1}^{n} w_i}}\right)$$

where I indicates the set of observations in hold-out set j.

We first explored the class of multiple linear regression models. The goal of these models was to predict the number of people living in each household as a function of all available covariates. Multiple linear regression models were fitted using the `svyglm` function in the `survey` package. See *Table 10* for a complete enumeration of models fit and their associated weighted RMSE. The best performing multiple linear regression model was that which included main effects for `BDSP`, `PUMA`, `unit_type`, and `include_tax`, as well as interactions between `include_tax` and `TAXAMT`, `include_tax` and `BDSP`, and `include_tax`, `TAXAMT` and `BDSP`. This model had a cross-validation weighted RMSE of 1.178. In context, this RMSE indicates that on average, our predicted number of people per household is off from the true number of people per household by 1.178 people.

We then explored the class of Poisson regression models. The goal of these models was to predict the number of people living in each household by modeling the bedroom occupancy rate, which we defined as the number of people per bedroom (`NP`/`BDSP`). However, because studios (zero-bedroom units) appeared in the ACS data, this created an issue. To account for this, we instead used the variable `BDSP_nostudio`, described above. Poisson models were also fitted using the `svyglm` package in `R`. In every case explored, the Poisson models performed less well than the best linear regression model. For example, the best-performing Poisson regression model was that which modeled the log occupancy rate as a function of `PUMA`, `unit_type`, and `include_tax`, as well as interactions between `include_tax` and `TAXAMT`, which resulted in a cross-validated RMSE of 1.215. *Table 11* listed each model and its performance.

Model Selection

Model performance was evaluated by comparing weighted root mean squared error estimates averaged across cross-validation folds (as described in 'Model Fitting'). The model which performed the best was a linear regression model using main effects for `BDSP`, `PUMA`, `unit_type`, and `include_tax`, as well as interactions between `include_tax` and `TAXAMT`, `include_tax` and `BDSP`, and `include_tax`, `TAXAMT` and `BDSP`. This model had a cross-validation weighted RMSE of 1.178. This model was thus selected as the model to be used for parcel-level predictions on the King County Assessor's data.

When refit using the entirety of the ACS training data, this model (hereafter referred to as "the best linear regression model") produced a weighted RMSE of 1.187 when used to predict on the ACS validation set. This model predicted 442,595 people for the ACS validation set, when in fact that sample represented 437,841 people, so at the county level this model overestimates by approximately 1.1%.

Prediction

We produced parcel-level predictions for every parcel in the King County Assessor's data using the best linear regression model, described above.

As an initial investigation of model performance, we summed our predictions for each parcel in King County and compared the resulting total to the known number of people living in King County. The sum of our predictions results in a county population of approximately 1.4 million people, compared to the true population of approximately 2.25 million people[19].

Our original plan was to compare parcel predictions aggregated to the Census tract-level to true tract populations, however it was clear after our initial investigation that such smaller scale estimates would not be accurate, given that aggregate estimates were so far off target.

Project Output Description

Our final deliverable is a model which predicts populations at the parcel level. We provided a detailed description of our modeling process to PHSKC, as well as a well-documented Github repository which recreates our modeling process. Further, we provided parcel-level population estimates for all parcels in the county.

Impact and Limitations

We provided Public Health Seattle and King County with a model which predicts parcel populations based on publicly available parcel data. These predictions and this modeling framework provides the county with a novel method to determine denominator populations for any geographical region of interest within the county, regardless of size or shape. Such population estimates were previously not possible due to constraints imposed by public data access. The potential uses of these denominator populations are nearly limitless. While PHSKC could use these estimates to track disease propagation or vaccination patterns in the county, the model or the estimates it produces could potentially be shared more broadly with other county departments. For instance, King County Local Services might be interested in using small area population estimates to determine where to plan future public transit expansions.

Despite our model's potential for future use, it does have some limitations. The primary limitation is that its accuracy is limited by the data we have available. The covariates which we could include in our model were limited to those which were made publicly available by the King County Assessor for every parcel in the county. While more accurate estimates could almost certainly be constructed if we had access to additional information about a parcel (such as household income, for example), such information was not available.

Furthermore, our outcome of interest, parcel population, is inherently unstable. As a result, while we might be able to make accurate estimates for a parcel of a given set of covariates on average, our estimates might vary from the truth somewhat dramatically for a single parcel. While this would likely not be an issue if the aggregate regions of interest were large enough, for very small aggregate regions (city blocks, for instance), such discrepancies might result in highly inaccurate estimates.

Additionally, we did not take group quarters into account in our model as survey weights for group housing were not included in the ACS data. According to the United Census Bureau, the total group quarters (GQ) population in the ACS might not be comparable with decennial census counts because some GQ types are out of scope in the ACS. For lower levels of geography, particularly when there were relatively few GQs in a geographic area, the ACS estimate of the GQ population would vary from the count from the decennial census suggesting that the use of GQs reported by the ACS may be inappropriate for making prediction. However, this problem would be solved if there were better available group housing data.

Finally, there is a clear discrepancy between model performance when the same model was used to predict on the ACS validation data and the Assessor's parcel data. We

suspect that such discrepancies are due in part to errors in the cleaning of the Assessor's data. These limitations should be kept in mind by the end data user.

Next Steps

An obvious next step for this project would be a proof of concept. The parcel populations predicted by our model could be aggregated to a region of interest and then used for an analysis by PHSKC. Ideally, the output of that analysis could be compared to the output produced by the traditional approach used by the county to construct denominator populations.

A possible extension of this project would be to consider how parcel level populations are changing over time, and to use this information to construct parcel level population estimates for both historical and future points in time. Such a project would be of interest to PHSKC for several purposes. First, it could allow the county to model how disease rates are changing over time and space (how the distribution of opioid use has shifted in Seattle between 2000 and today, for instance). It could also allow the county to plan resource allocation for the future (which parts of the county might benefit the most from vaccine advocacy programs in public schools, for instance).

Even in the absence of a model which predicts future parcel change, our model could be used in future years on newly released Assessor's data in order to construct current-day parcel estimates. That is, as public tax information is released on a yearly basis, that information could be fed into our current modeling framework to construct up-to-date parcel population estimates. While this wouldn't allow for prediction into the future, after several years of doing this process, a database of historical population estimates would exist which could be used by PHSKC for various purposes.

Output Transition Plan to Sponsor

The output transition plan to our sponsor, Daniel Casey of PHSKC, consists of three stages. The initial step was to present our findings to Daniel and other interested parties on March 4th, 2024. This allowed Daniel to provide final feedback on our project and ensure the output meets his requirements and that of PHSKC.

Upon completion of the presentation and any final feedback, we will provide Daniel with the final report in PDF or Word format, depending on his preference, as well as our presentation via email. Should he have any other requests regarding dissemination of our methods or findings we can incorporate into this stage as well.

Finally, we will document our progress on a shared GitHub project repository, ensuring not only reproducibility of our findings, but also the ability to transfer our codes and results to PHSKC for future use and research. Daniel will be granted access to this page as an administrator so that full rights and permissions will be granted should he plan to employ our research.

Team Member Contributions

For draft 1, Lucas copied relevant sections from our SAP and updated tense, Gavin wrote the Abstract and Output Transition Plan to Sponsor and produced descriptive statistics for the Assessor's data, and Erika wrote Impact and Limitations, Next Steps, and the Approach/Methods overview and produced descriptive statistics for the ACS data.
For draft 2, Erika incorporated Lloyd's feedback from draft 1 and expanded the Methods section, Gavin discussed the Assessor data cleaning and the random forest model, and Lucas explored alternative models and updated citations.
For the final draft, Erika implemented Lloyd's suggestions from draft 2 and incorporated final results from our modeling process.

ACS DATA

Table 1: Counts by Public Use Microdata Area, non-group housing

PUMA	Unweighted Count	Weighted Count
11601	628	78567
11602	557	56453
11603	591	94339
11604	474	59836
11605	604	62196
11606	529	51703
11607	608	65951
11608	673	62375
11609	546	58372
11610	463	56225
11611	500	49466
11612	424	49100
11613	437	45780
11614	456	44286
11615	499	45927
11616	464	43406

Table 2: Counts by unit type, non-group housing

Unit Type	Unweighted Count	Weighted Count
Apartment	2601	364628
Condo	470	53550
House	5382	505804

Table 4: Descriptive statistics, non-group housing

Variable	Weighted Count	Mean	SD	Minimum	Median	Maximum
People	923982	2.377636	1.383709	1	2	14
Bedrooms	923982	2.574650	1.385936	0	3	7
Tax amount ($)	525122	7003.490808	5401.227996	0	6500	32500

ASSESSOR'S OFFICE DATA

Table 5: Count by unit type - Assessor's Office data

Unit Type	Documented Count
Apartment	9862
Condo	75650
House	520205

Table 6: Descriptive statistics, bed data - Assessor's Office data

Variable	Mean	SD	Median	Minimum	Maximum
Bed Count	3.36	5.85	3	0	633

Table 7: Descriptive statistics, unadjusted housing lots - Assessor's Office data

Variable	Mean	SD	Minimum	Median	Maximum
Lot Size (Sq Ft)	72764.34	1029150	0	8025	37932048

Table 8: Count by acreage type - Assessor's Office data

Acreage	Adjusted Count
Less than 1 acre	522534
Between 1 and 10 acres	42683
Greater than 10 acres	6602

Table 10: Results of Linear Regression Models

Model	Weighted RMSE
$E[NP] = BDSP + PUMA + include\_tax + include\_tax * TAXAMT$	1.187
$E[NP] = BDSP + unit\_type + include\_tax + include\_tax * TAXAMT$	1.191
$E[NP] = PUMA + unit\_type + include\_tax + include\_tax * TAXAMT$	1.265
$E[NP] = BDSP + PUMA + unit\_type$	1.199
$E[NP] = BDSP + PUMA + include\_tax + include\_tax * TAXAMT + unit\_type$	1.184
$E[NP] = BDSP + PUMA + BDSP * PUMA + include\_tax + include\_tax * TAXAMT + unit\_type$	1.187
$E[NP] = BDSP + PUMA + BDSP * unit_type + include\_tax + include\_tax * TAXAMT + unit\_type$	1.184
$E[NP] = BDSP + PUMA + include\_tax + include\_tax * TAXAMT + unit\_type + BDSP * include\_tax + BDSP * include\_tax * TAXAMT$	1.178
$E[NP] = BDSP + PUMA + BDSP * PUMA + include\_tax + include\_tax * TAXAMT + BDSP * unit\_type$	1.187

Table 11: Results of Poisson Regression Models

Model	Weighted RMSE
$log(E[NP]) = log(BDSP\_nostudio) + PUMA + include\_tax + include\_tax * TAXAMT$	1.225
$log(E[NP]) = log(BDSP\_nostudio) + unit\_type + include\_tax + include\_tax * TAXAMT$	1.22
$log(E[NP]) = log(BDSP\_nostudio) + PUMA + unit\_type$	1.245
$log(E[NP]) = log(BDSP\_nostudio) + PUMA + include\_tax + include\_tax * TAXAMT + unit\_type$	1.215
$log(E[NP]) = log(BDSP\_nostudio) + PUMA + include\_tax + include\_tax * TAXAMT + unit\_type + PUMA * include\_tax + PUMA * include\_tax * TAXAMT$	1.22
$log(E[NP]) = log(BDSP\_nostudio) + PUMA + include\_tax + include\_tax * TAXAMT + unit\_type + PUMA * unit_type$	1.22

References

1. About Public Health - Seattle King County. Accessed February 1, 2024. https://kingcounty.gov/en/dept/dph/about-king-county/about-public-health/administration
2. Disclosure Avoidance for the 2020 Census: An Introduction. US Department of Congress. US Census Bureau. November 2021. https://www2.census.gov/library/publications/decennial/2020/2020-census-disclosure-avoidance-handbook.pdf
3. Glossary. United States Census Bureau. Accessed Nov 6, 2023. https://www.census.gov/programs-surveys/geography/about/glossary.html#par_textimage_13
4. Parcel Definition. Law Insider. Accessed Nov 6, 2023. https://www.lawinsider.com/dictionary/parcel
5. Assessments Download Data, King County Department of Assessments. Accessed Nov 6, 2023. https://info.kingcounty.gov/assessor/datadownload/default.aspx
6. Look up Property Info. King County. Accessed Nov 6, 2023. https://kingcounty.gov/en/legacy/depts/assessor/parcel-sales-search
7. American Community Survey. United States Census Bureau. Accessed Nov 6, 2023. https://www.census.gov/programs-surveys/acs/about.html
8. United States Census Bureau. Accessed Dec 12, 2023. https://www2.census.gov/geo/maps/DC2020/PUMA/st53_wa/
9. Public Use Microdata Areas (PUMAs). United States Census Bureau. Accessed Dec 12, 2023. https://www.census.gov/programs-surveys/geography/guidance/geo-areas/pumas.html
10. ACS 5-Year Estimates Public Use Microdata Sample (2021). United States Census Bureau Beta. Accessed Dec 12, 2023.
11. Decennial Census of Populations. United States Census Bureau. Accessed Nov 6, 2023. https://www.census.gov/programs-surveys/decennial-census.html
12. 2020 Census Public Use Microdata Area (PUMA) Reference Maps. United States Census Bureau. Accessed Nov 20, 2023. https://www2.census.gov/geo/maps/DC2020/PUMA/st53_wa/DC20PUMA_5323301.pdf
13. Parcel Viewer. King County. Accessed Nov 20, 2023. https://gismaps.kingcounty.gov/parcelviewer2/
14. Seattle Zip Code Map. US Map Guide. Accessed Nov 20, 2023. https://www.usmapguide.com/washington/seattle-zip-code-map/
15. 2020 Census Tracts Seattle. Seattle GeoData. Accessed Nov 20, 2023. https://data-seattlecitygis.opendata.arcgis.com/datasets/9075e8c912a24c4b9458af8866c72ae7/explore?location=47.596902%2C-122.252932%2C11.00
16. American Community Survey and Puerto Rico Community Survey Design and Methodology. Version 3.0. United States Census Bureau. Issued November 2022.

17. Understanding and Using American Community Survey Data: What All Data Users Need to Know. United States Census Bureau. Accessed Feb 20, 2024. https://www.census.gov/content/dam/Census/library/publications/2020/acs/acs_general_handbook_2020.pdf
18. IPUMS USA. Accessed March 10, 2024. https://usa.ipums.org/usa/about.shtml
19. Demographics. King County. Accessed March 10, 2024. https://kingcounty.gov/en/dept/executive/governance-leadership/performance-strategy-budget/regional-planning/demographics