

Survey Address Selection

1 Preamble

1.1 Contractor: Joel Adria, 780-437-8113, joel.adria@gmail.com

1.2 Description: Obtain 5000 random Edmonton names and addresses in a database or spreadsheet, as a distribution list for surveys.

1.3 Approach: Using known telephone exchange information and random generation of phone numbers targeted at the reverse lookup features of Canada411.com to obtain random names and addresses.

2 Technical Details

2.1 Exchange Selection

By using only Edmonton-based exchanges, which is the first 3 numbers after the area code, phone numbers from only the Edmonton area can be generated. By using an online resource (<http://telcodata.us>) a list of Edmonton exchanges was obtained.

The following exchanges were identified as being within Edmonton: '341', '342', '371', '377', '378', '391', '392', '395', '399', '401', '405', '406', '407', '408', '409', '412', '413', '414', '415', '420', '421', '422', '423', '424', '425', '426', '427', '428', '429', '430', '431', '432', '433', '434', '435', '436', '437', '438', '439', '440', '441', '442', '443', '444', '445', '446', '447', '448', '450', '451', '452', '453', '454', '455', '456', '457', '461', '462', '463', '465', '466', '468', '469', '471', '472', '473', '474', '475', '476', '477', '478', '479', '480', '481', '482', '483', '484', '485', '486', '487', '488', '489', '490', '491', '492', '493', '495', '496', '497', '498', '499'.

From this list, the program randomly selected one of these first three numbers.

2.2 Suffix Selection

Four more digits from 0 to 9 were then generated, one at a time. The source of this random selection was the built-in random generating facilities inside the programming language used, Python. While these facilities are not truly random (the numbers are being generated using pseudo-random, or from an algorithm), this was not found to be a significant issue, as phone numbers are chosen at random and are not associative with the selection criteria: location, income, etc.

2.3 Data Query

Once this number is generated, it is then submitted to the reverse phone lookup form of Canada411.com Person Search. The method for accessing this form was conveniently available in GET format, therefore lookup was simply a matter of modifying the URL being requested with the appropriate area code, exchange, and suffix parameters.

2.4 Valid Result Detection

Once the web page data is returned to the script, it is scanned at line 168 for the presence of the words "No listings", which would cause the program to skip the result and try another number. Approximately 75% of these numbers did not have listing, either because they are not in service, or because they are not listed in the phone book.

If the previous test fails, then valid results are therefore present, and another scan is done to locate the presence of the word “Edmonton” on line 355 of the data. In the event that the result was a business, the layout changes and the word “Edmonton” is no longer on line 355, therefore this test eliminates non-personal results. Also, if the address is not listed with the name and address, the words “Edmonton” are not listed, so it too is thrown out.

If both the listing and location tests are passed, the program saves the page in its entirety to a folder for archival, and later processing.

A counter keeps track of how many successful results have been obtained, and once the counter has been fulfilled the program stops. To speed up the process, three instances of the script were run in parallel, after which the were combined.

2.5 Further Processing

After 5000 web pages of information had been archived, another script was created to process these pages. While it would have been slightly more efficient to process the data when they were being retrieved, this method provided a reliable way to review results in the event that the first program did not complete successfully and lost the results.

This second script looped through all the web page files in the directory and read and separated the data into values, and then stored these values in a CSV file for opening in any spreadsheet application.

The information in the spreadsheet was very reliable, with a few results that did not contain complete addresses or postal codes due to inaccuracies from Canada411.com, and these results were removed and new results generated to replace them. In addition, the organizer of the project, Dr. Marco Adria, was coincidentally chosen in the process and was also removed. Some postal codes were not supplied from Canada411.com and had to be added manually. It was recommended that these results not simply be removed and replaced, as the majority of them had the common trait of named boulevards, streets, which could be associated with certain higher income neighborhoods, etc. and removing them could skew the results.

2.6 Visualization

While the randomness of this process was reasonably correct and the spreadsheet appeared to be from different locations around the city, a more useful way to visualize this would be to use address information and display a map of the results. To solve this problem, all addresses were processed using tools from <http://www.gpsvisualizer.com/geocoder/> that took advantage of Yahoo! address to GPS coordinate conversion abilities. The same website provided conversion tools to output to KML, a language used to display overlay objects in Google Earth. After batch-processing five groups of 1000, the data was merged to visually represent the participants. (*See Figure 2.1*)

The results were impressive. Participants were adequately selected from all parts of the city, including smaller more remote areas that are still considered parts of Edmonton. The uncovered bands to the south-east and north-west are industrial areas, and therefore did not yield any results, as desired.

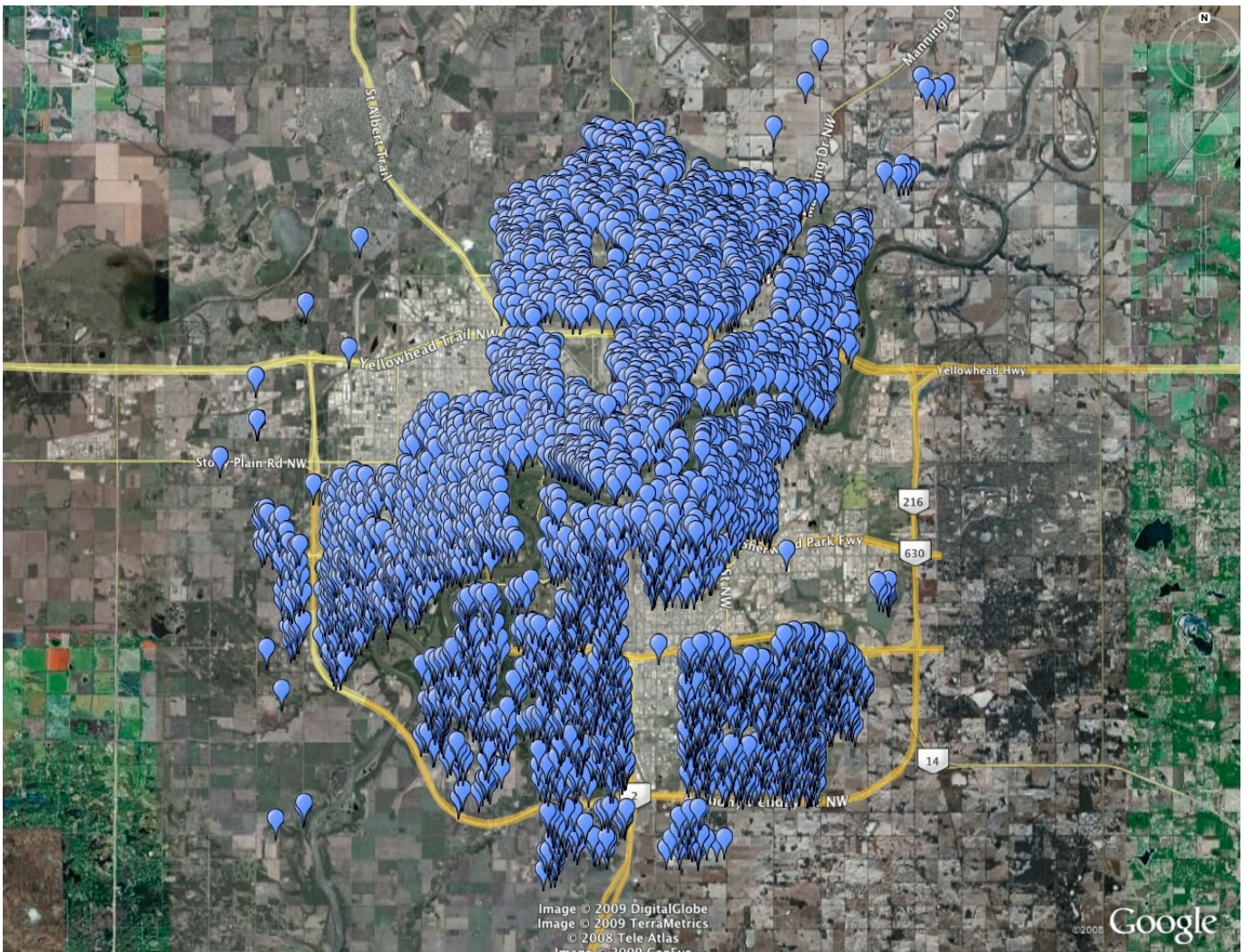


Figure 1.1 : Edmonton map with participants represented by markers in Google Earth.