Data Transformation and Geocoding of Historical City Directories

Eva Dodsworth[*1] and Markus Wieland[†1]

[1]Geospatial Centre, University of Waterloo Library

January 7, 2022

Summary

This paper explores the geocoding methods used on historical city directory addresses for the purposes of studying and visualizing 100 years of urban development, business growth, as well as streetscape changes for a small city in Ontario, Canada. Using historical city directories, alongside maps and fire insurance plans, the researchers studied cardinal changes, street name changes and address number shifts for the years 1900-2000 to create a seamless open-access discoverable and interactive map. Users can look up a century's worth of business information for endless research opportunities, as well as study families and their movements throughout the decades.

KEYWORDS: city directories, historical GIS, street network, geocoding, demography

1. Introduction

City directories offer extraordinary insight into what cities and communities were like at a given point in time, providing researchers, genealogists, property owners, and the general public with a history of household residents, their occupations, business information, and locations of churches, schools, courts, and much more (Basar, 1999).

This project, that we coined *Urbanizing Kitchener: A Historical Geospatial Perspective* promotes quantitative historical research and facilitates visualization by 1) providing researchers with an organized online map and database listing one-hundred years' worth of city directory entries, and 2) providing users with a historical narrative related to local residential and commercial urbanization. The online map organizes the directory entries spatially, by historical and modern street addresses, and enables query and location specific search criteria, inviting researchers to conduct spatial analysis, gain new insights, ask new questions based on geographic proximities, and transform their research and teaching as it relates to urban structure and population geography.

U.S. Researchers have mapped city directory entries in the past, focusing on finding specific patterns related to restaurant growth (Brackhan, 2009), discovering ethnic patterns (Hardwick, 1990), and identifying brownfield properties (Hayek, et al, 2010). This project aims to analyze and answer questions around the evolution of businesses and establishments in Kitchener, Ontario between the years 1900-2000, examining the types of businesses, their migration patterns, adaptive reuses, as well as urban regeneration. The study of residential spaces will provide insight on family clusters, workplace locations, occupational mobility, and distances to work.

This project consists of three distinct stages: 1) the creation of geospatial data: taking the entries from the city directories, uncluttering and converting historical addresses to modern ones; 2) developing a user-friendly online mapping program and importing the geospatial data collection into it; 3) conducting research and spatial analysis, and presenting results for purposes of transferring knowledge, and promoting the application to other researchers and community users.

[*] edodsworth@uwaterloo.ca
[†] markus.wieland@uwaterloo.ca

2. Methodology

2.1 Data Collection

The initial step of acquiring the original data came in the form of scanned PDF files of the Vernon directories (1900-2000) from the local Kitchener Public Library. The process of scanning documents using Optical Character Recognition (OCR) picks up many artifacts that need to be filtered out when transcribing from PDF to CSV. Some examples of OCR issues include uppercase W's displayed as AA7, AV, A\, AAr, as well as common artifacts displayed throughout the documents, such as • (^ * ? ! | ° } _ < ;.

The earliest directories had many artifacts, that coupled with changes to layout made attempts at automating transcription extremely difficult. Transposing the directories to CSV's, line by line was the best way to utilise the varied skills people had who were involved in this endeavour. The quality of the OCR scans improved in later years and transcribing has switched from line by line to a comparison model between years, only modifying, adding or removing changed rows.

2.2 QA Stages

The next four stages of QA form the backbone of the project and consist of a combination of visual and programmatic processes. These stages ensure the datasets are optimised for further processing and storage.

Stage 1: Review transcribed CSV's to ensure that data is in the correct columns using Excel formulas, table filtering and find and replace. Columns include Surname, First Name, Occupation, and Historical Address

Stage 2: Levenshtein Distance Stage 1:
This is a python script that checks abbreviations of certain words and updates them to full words. Eg. Saw to Sawyer or Sec to Secretary. The directory uses acronyms and short forms for occupations so this step updates old words to a modern equivalent and/or changes the abbreviation to the full word for easier reading.

Stage 3:
This review pays particular attention to the Historical Address column. This is to ensure that the data cell format is Number, Street Name, Street abbreviation, and cardinal direction for the next stage.

Stage 4: Levenshtein Distance Stage 2:
Cities change quite a bit over the span of a hundred years, including address number shifts and street name changes. Historical addresses need to be geolocated to the correct position on a modern map. This process is a little bit more involved than the Stage 2 Levenshtein Distance because the address dataset needs to be broken up into its constituent components (Number, Street Name, Street Abbreviation, and Cardinal Direction). The Levenshtein Distance algorithm in the python code is only used on the street name section of the string in the historical address column cells. The rest of the string must be parsed and reassembled with either the same street name (if not changed) or the updated street name (if it is changed). Along with the Levenshtein Distance algorithm there are many checks and changes in this code to assemble the correct address for the geocoder. One check adds a flag (OANF – Original Address Not Found) in the Note column. This flag states that the modern equivalent address doesn't exist in a geo-codable place and pulls a manually created X and Y Coordinate from the updater/lookup CSV. These flags were created to automate the avoidance of mismatches in the geocoding process.

Since there are going to be millions of rows of data in the directory database they all must be prepared for geocoding using the Levenstein ratio. This finds, replaces, and corrects any addresses that may be misspelled or similar.

Table 1 and **Table 2** are examples of cleaned data prepared for database ingestion.

Table 1 Business Dataset

Business Name	Business Type	Historical Address	Current Address	Note	X	Y	Images
5 Point Meats	Grocery	204 Frederick	214 Frederick	OANF	-80.4818865	-43.45388907	Image File loc.

Table 2 Residential Dataset

Last Name	First Name	OCCUPATION	Historical Address	Current Address	Note	X	Y
Moser	Geo	Proprietor Hotel Brunswick	101 King W	106 King W	OANF	-80.49108664	-43.45063406

2.3 Geocoding

To increase precision and repeatability we built our own composite address locater in Esri Arc Desktop, comprising of three shapefiles. The first is the most accurate made from the city's address point file. The second are address ranges (left and right side) based on the city's road network. Lastly, the third is the city's road line (street names) shapefile. This composite locator catches all the possible variations in addresses and can be modified to emulate address ranges that existed during the years being geocoded.

2.4 Web Map

The presentation and interactive component is made up of Postgres database backend, React, NodeJS and Ingnx middleware, and leaflet API webpage. The final CSV's created from the previous processes are loaded into the Postgres database using React based administration page. The X and Y coordinates are interpolated by the Leaflet API creating the geolocated address markers on the web map. These markers have descriptors from the rows in the database that are also represented in a table below the map. The site also has trigram-based search functionality to search by name, historic address, current address, occupation, business name, business type and year or any combination of these.

3. RESULTS

This online data transformation project isn't only an effort to digitize and make historical documents more easily accessible, but rather this initiative significantly improves and enhances the original in-print Vernon city directory usability experience by offering access to all directory years, searchable by residential or business address, resident's last name, occupation, place of employment, as well as business type, business name and even business advertisement. Wherever possible, spelling errors were corrected, duplicate entries were removed and most importantly, street address name changes and house number shifts have been meticulously researched, captured, highlighted and enhanced with modern equivalents. Therefore, when, for example the current location of a historical record may be in the middle of an intersection, the project's goal is to keep that historical fact known and traceable. Whether demolished and replaced with a wider road, or an apartment building, or whether the original building is still standing, *Urbanizing Kitchener: A Historical Geospatial Perspective* will give users the tools to learn more about each address, its occupants and their journeys from 1900 to 1950 and eventually beyond. **Figure 1** demonstrates the spatial benefits of geocoding business addresses by presenting the locations of all businesses and highlighting the geographic areas that had high density distributions.

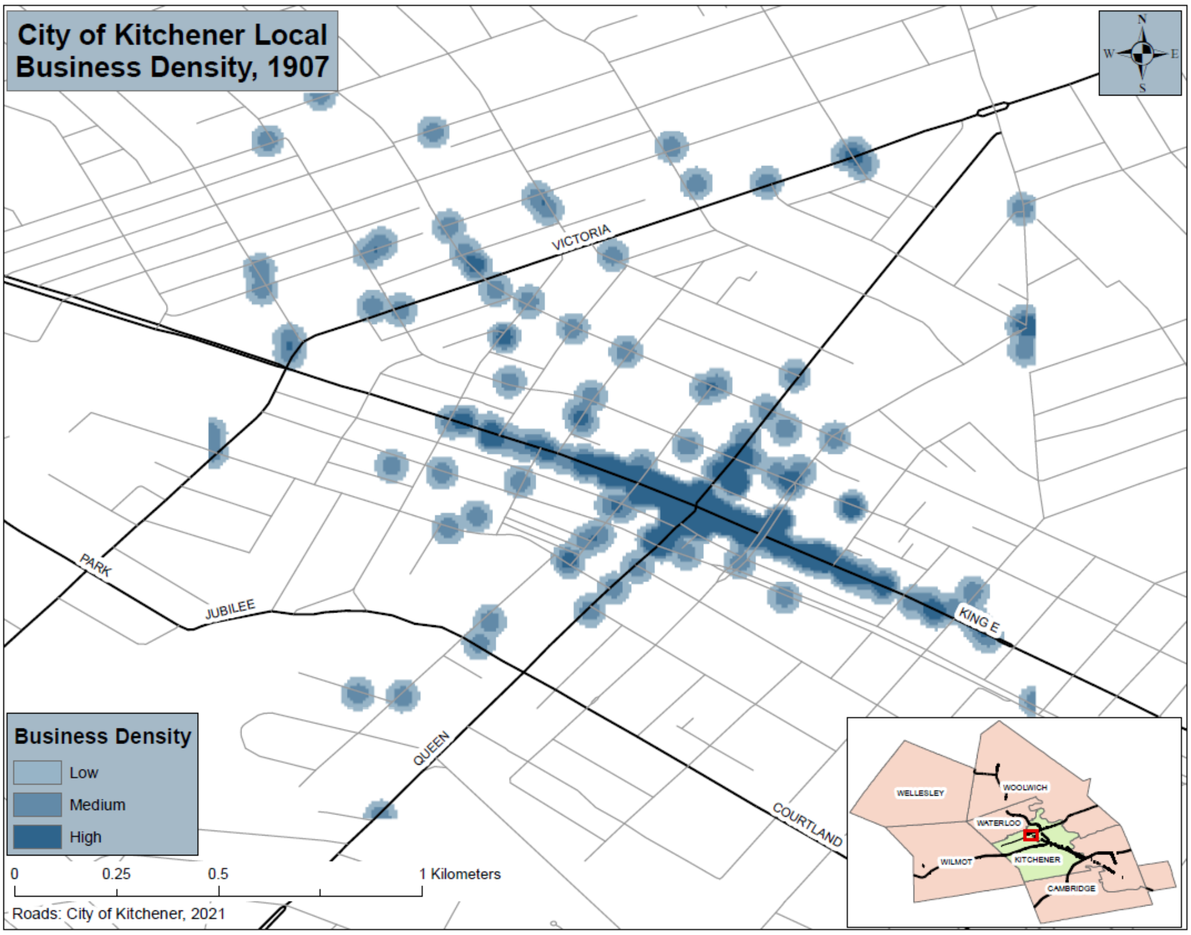

Figure 1 Business Density in Kitchener, Ontario in 1907

The searchable interactive online map organizes the advertisements, household and business entries by street address, enabling query and location specific searches, allowing researchers to conduct spatial analysis, gain new insights, ask new questions based on geographic proximities, and transform their research and teaching as it relates to urban structure and population geography.

4. Discussion

Urbanizing Kitchener: A Historical Geospatial Perspective is a project that modernizes ‘old data’, and not only offers a primary resource to newer scholars who may be unaware of the resource’s existence, but offers it in a convenient and contemporary way. As is often the case with historical documents, when resources are difficult to access, they may be deemed obsolescent and simply are overlooked. Taking historical documents and making them easily accessible not only breaths life back into once-deemed rich resources, but it moves it from stagnancy into active use. This project specifically uses geo-technologies to reveal pertinent patterns that may otherwise be missed if not spatially presented. This is the first project of its kind to geocode and map one hundred years of city directory information, and the researchers plan to share the documentation and results with others in hopes to have others build on this project for other cities in the future.

One unexpected outcome of this project is the creation of a historical-to-modern address lookup resource. The city of Kitchener does not have historical records of street changes before 1950, so this project inadvertently produced a valuable historical resource as well.

5. Acknowledgements

This project has been funded in part by the Social Sciences and Humanities Research Council, as well as the University of Waterloo. We would like to acknowledge support from Karen Ball-Pyatt, from the Kitchener Public Library, Darryl Bonk, from Waterloo Region Generations, Waterloo Regional Heritage Foundation, and the Waterloo Region Branch Ontario Genealogical Society who have generously provided us with scanned versions of the city directories.

References

Basar, I. (1999). Directory publishing in Canada: the last hundred years. *The Serials Librarian*, *37*(1), 59-82.

Brackhan, J. L. (2009). *Restaurant growth in Lawrence, Kansas, 1950 to 2007* (Doctoral dissertation, University of Kansas).

Educative (2021). The Levenshtein distance algorithm. https://www.educative.io/edpresso/the-levenshtein-distance-algorithm.

Hardwick, S. W. (1990). Using city directories to teach geography. *Journal of Geography*, *89*(6), 266-271.

Vernon's Berlin, Waterloo and Bridgeport Directory (1901-1913). Hamilton, Ontario: H. Vernon.

Vernon's Kitchener-Waterloo city directories (1919-1947). Hamilton, Ontario: H. Vernon.

Biographies

Eva Dodsworth is the geospatial data services librarian at the University of Waterloo library where she specializes in teaching GIS and map-related content to the university community. Eva's interests include historical cartographic research, teaching geoweb applications and historical GIS. Eva is also a GIS instructor for a number of academic institutions.

Markus Wieland is the GIS specialist at the University of Waterloo library where his guidance has helped researchers change environmental policy, Canada's building code and with papers and publications. He started as a software engineer, moved on to technical design and finally GIS where he has been an educator, technician and now specialist.