9-21-2022

UNI·CEN Documentation Report 3: Digital Boundary Files

Zack Taylor
Dept. of Political Science, Western University, zack.taylor@uwo.ca

Christopher Macdonald Hewitt
Network for Economic and Social Trends, Western University, chewitt6@uwo.ca

unicen

Unified Infrastructure for Canadian Census Research

Documentation Report 3

Digital Boundary Files

Dr. Zack Taylor
Dr. Christopher Macdonald Hewitt

September 21, 2022
Version 1.0

Western SocialScience | Network for Economic and Social Trends (NEST)

Contents

The UNI·CEN Project

Analysis of historical and contemporary Census data is an active area of research. Sociologists, historians, geographers, urban and regional planners, and political scientists have used these data to study the historical development of, and change in, international and domestic migration, urban settlement patterns, inter-group relations, economic change, and political representation. In the United States, United Kingdom, and other countries, projects increased the accessibility of historical Census data by compiling existing digital datasets, digitizing those that exist only in print, and creating modern systems to disseminate them to users.

The **UNI·CEN** (Unified Infrastructure for Canadian Census Research) project follows the example of these international projects. We compiled available aggregate Census data at several commonly used levels of geography for the 1851–2021 period and converted it to a standardized table format. We digitized mapped boundaries, data tables, and geographic coding schemes pertaining to census tracts for the 1951–66 period. We developed a standardized variable naming system and coded the compiled data with it, enabling analysis and visualization of change over time. We also developed a set of geographic linkage tables that enable comparison of places across time despite inconsistent naming and coding. Finally, we have assembled available corresponding digital boundary files and reformatted them to join to the data.

Undertaken between 2018 and 2022, **UNI·CEN** is a project of Western University's **Network for Economic and Social Trends** (NEST). **UNI·CEN** parallels a companion project, the **Canadian Communities Policy Observatory** (https://observatory.uwo.ca), a portal that enables visualization, analysis, and retrieval of place-based data. Both projects are funded by Western University's Faculty of Social Science.

Project Team

Investigators, UNI·CEN Project

- Dr. Zack Taylor, Project Leader and Associate Professor, Department of Political Science, Western University
- Dr. Victoria Esses, Professor, Department of Psychology, Western University
- Dr. Dave Armstrong, Associate Professor, Department of Political Science, Western University

Postdoctoral Fellow (Mitacs – Esri Canada)

- Dr. Christopher Macdonald Hewitt, Western University

Overview

The **UNI·CEN Digital Boundary Files** series contain reformatted versions of all publicly available digital boundary files. The original files come from a variety of sources created at different times and for different purposes. As a result, they are stored in a variety of file formats, use different projections, and their attribute tables and metadata contain different items. They also handle shorelines and internal water bodies differently.

The goal of the this project is to create a series of digital boundaries that:

- Standardize attribute table fields
- Make the handling of shorelines and internal water bodies as consistent as possible
- Are projected to the same coordinate system
- Are simplified to a similar level of detail
- Are stored in modern file formats

File series

Three series of boundary files are created, differentiated by their treatment of shorelines and internal water bodies.

1. **CBF-Harmonized Shoreline series**: Intended for cartographic use, these files have been clipped to a standardized ocean and Great Lakes shoreline and, with the exception of 1986 where necessary source materials do not exist, have no internal water bodies.
2. **CBF-Original Shoreline series:** These files maintain the shorelines found in the original files that are part of Statistics Canada's Cartographic Boundary Files (CBF) series, or which are derived from them (i.e., those from The Canadian Peoples Project, the Canadian Century Research Infrastructure project, as well as census tract digitization projects undertaken by Christopher Hewitt and ScholarsGeoportal).
3. **DBF-Original Shoreline series:** These files maintain the shorelines found in Statistics Canada's Digital Boundary Files (DBF), which have highly generalized coastal boundaries that extend into water bodies.

Source CBFs and DBFs are not available for all years. For Statistics Canada's description of the distinction between CBFs and DBFs, visit: https://www12.statcan.gc.ca/census-recensement/2021/geo/sip-pis/boundary-limites/index2021-eng.cfm?year=21.

Those interested in using these files for detailed cartographic representation or analysis should supplement them with Statistics Canada's Coastal Waters and Lakes and Rivers files.[1]

[1] https://www12.statcan.gc.ca/census-recensement/2011/geo/bound-limit/bound-limit-2006-eng.cfm

Dissemination location

UNI·CEN Digital Boundary Files are stored on the open-access Borealis Dataverse repository at: https://borealisdata.ca/dataverse/unicen_boundaries. As of this publication, only the **CBF-Harmonized Shoreline** series has been made available.

Disclaimer

We have made every effort to verify the accuracy of the boundary files and datasets produced by this project. We ask that users notify us of any errors they discover so that they can be corrected in future versions.

Acknowledgments

We are grateful for the assistance of Dr. Jeff Allen, School of Cities, University of Toronto; Amber Leahey, Data and GIS Librarian, Scholars Portal; and Leanne Trimble, Data Librarian, University of Toronto.

Levels of geography

Boundary files were standardized for five levels of geographic aggregation:

- census tract (CT)
- census metropolitan area/census agglomeration (CMA)
- census subdivision (CSD)
- census division (CD)
- province/territory (PR)

At various points in time, Statistics Canada and predecessor organizations have disseminated census data at other levels of geographic aggregation, for example, enumeration areas, dissemination areas, and economic regions. In restricting the project to the levels listed, our goal was to privilege levels of geographic aggregation for which data have been disseminated for extended periods of time, and which are most commonly used by researchers.

Federal electoral districts (FED) will be added in the future.

Temporal coverage

Not all geographic levels are available in all years. The PR, CD, and CSD levels of geography (and equivalents in earlier years) are available for most census years between 1851 and 2021, with the following exceptions:

- CSD and CD boundaries for 1956, 1961, 1966, 1971, and 1976 have never been digitized.
- PR boundaries for these years have also never been released digitally, however given that provincial boundaries did not change during this period we plan to create these from later-year files.

The CT and CMA levels of geography were first disseminated in 1951. After digitizing the 1950s- and 1960s-era tracts from printed maps (see Report 4 for a description of this project), and reconstructing the 1971 boundaries from an alternative source, the series contains a complete set of CT and CMA boundaries for the 1951–2021 period.

Data sources

Table 1 summarizes the sources of the files used, indicating the availability of cartographic boundary files (CBF) and digital boundary files (DBF) as described above, and also the file source.

Table 1: Data Source Summary

Year	CBF	DBF	Source
1851	CSD, CD, PR		TCP
1861	CSD, CD, PR		TCP
1871	CSD, CD, PR		TCP
1881	CSD, CD, PR		TCP
1891	CSD, CD, PR		TCP
1901	CSD, CD, PR		TCP
1911	CSD, CD, PR		TCP
1921	CSD, CD, PR		TCP
1931	CSD, CD, PR		CCRI
1941	CSD, CD, PR		CCRI
1951	CSD, CD, PR, CT, CMA		CSD, CD, PR – CCRI / CT, CMA – Hewitt
1956	CT, CMA, PR		Hewitt
1961	CT, CMA, PR		Hewitt
1966	CT, CMA, PR		Hewitt
1971	CT, CMA, PR		CT, CMA – Allen / PR – Hewitt
1976	CT, CMA, PR		CT, CMA – SGP / PR – Hewitt
1981	CT, CMA	CSD, CD, PR	SC
1986	CSD, CD, PR, CT, CMA		SC except CT – Allen
1991	CSD, CD, PR, CT, CMA	CSD, CD, PR, CT, CMA	SC except CT – Allen
1996	CSD, CD, PR, CT, CMA	CSD, CD, PR, CT, CMA	SC except CT – Allen
2001	CSD, CD, PR, CT, CMA	CSD, CD, PR, CT, CMA	SC
2006	CSD, CD, PR, CT, CMA	CSD, CD, PR, CT, CMA	SC
2011	CSD, CD, PR, CT, CMA	CSD, CD, PR, CT, CMA	SC
2016	CSD, CD, PR, CT, CMA	CSD, CD, PR, CT, CMA	SC
2021	CSD, CD, PR, CT, CMA	CSD, CD, PR, CT, CMA	SC

Sources: **TCP** = The Canadian Peoples Project; **CCRI** = Canadian Century Research Infrastructure; **Hewitt** = Digitized by Christopher Hewitt, Western University; **Allen** = Rectified by Jeff Allen, School of Cities, University of Toronto; **SGP** = Digitized by ScholarsGeoportal; **SC** = Statistics Canada.

CSD source boundary files for the 1851–1921 period were digitized from paper references by The Canadian Peoples (TCP) project. For more information, visit https://thecanadianpeoples.com. CD and PR boundaries were created by aggregating CSD boundaries.

CSD and CD boundaries for 1931, 1941, and 1951 were digitized by the Canadian Century Research Infrastructure (CCRI) project. For more information, visit https://ccri.library.ualberta.ca/enindex.html. The publicly available CSD and CD shapefiles removed selected internal lakes and rivers. For greater comparability with boundaries in other years' files, we obtained the original geodatabases used to create the shapefiles from Byron Moldofsky, formerly of the University of Toronto's cartography. We extracted the polygon layers that did not have the lakes and rivers removed and used them as our input files. PR boundaries were created by aggregating CSD boundaries.

CT and CMA boundaries for 1951, 1956, 1961, and 1966 were digitized from paper references by Christopher Hewitt as part of a project led by Zack Taylor at Western University. The procedures used to create those files are described in Report 4.

PR boundaries for 1956, 1961, 1966, 1971, and 1976 are copies of the 1996 PR boundary file. Provincial and territorial boundaries did not change between Newfoundland Confederation in 1949 and the creation of Nunavut in 1999.

CT boundaries for 1971 exist on ScholarsGeoportal, but are highly generalized (see https://geo.scholarsportal.info/#r/details/_uri@=1042838780). Jeff Allen generated a new set of 1971 CT boundaries from Statistics Canada's 1971 Enumeration Area Reconstruction Project (see https://www150.statcan.gc.ca/n1/pub/16-510-x/16-510-x2017001-eng.htm). The smaller EAs were aggregated to CTs and compared to the paper maps and existing CT file. Discrepancies were rectified to match the paper maps. The EAs covered all CMAs except for Sarnia and Sault Ste. Marie. For these two regions, CT boundaries were copied from 1976 and then adjusted with reference to 1971 paper maps. CMA boundaries were created by aggregating CT boundaries.

CT boundaries for 1976 were digitized by ScholarsGeoportal as an internal project. While not yet publicly available, SGP generously shared the boundaries with us. CMA boundaries were created by aggregating CT boundaries.

Boundaries for the 1981–2021 period for all levels of geography created by Statistics Canada are available from ScholarsGeoPortal. The 1986, 1991, and 1996 CT files have been modified to correct non-systematic spatial mismatch problems. The rectification process used in 1991 and 1996, which involved a conflation procedure, is described in Allen, J. and Taylor, Z. (2018), "A new tool for neighbourhood change research: The Canadian Longitudinal Census Tract Database, 1971–2016." *The Canadian Geographer / Le Géographe canadien* 62: 575–58, https://doi.org/10.1111/cag.12467. The 1986 file was modified to address a systematic spatial mismatch with the Ottawa CMA only; non-systematic mismatches in the other CMAs have not

yet been corrected. The available 1991 CSD DBF file contained a large number of topology errors. Instead of correcting them, a new CSD file was created by aggregating enumeration area polygons available from ScholarsGeoPortal.

Source file processing procedure

After acquiring the source files, we reformatted them using the following procedure, which is discussed in detail below:

1. Correct topology errors
2. Simplify polygons
3. *For the CBF-Harmonized Shorelines series only:* Clip polygons to a coastal boundaries polygon.
4. Standardize attribute table
5. Export files

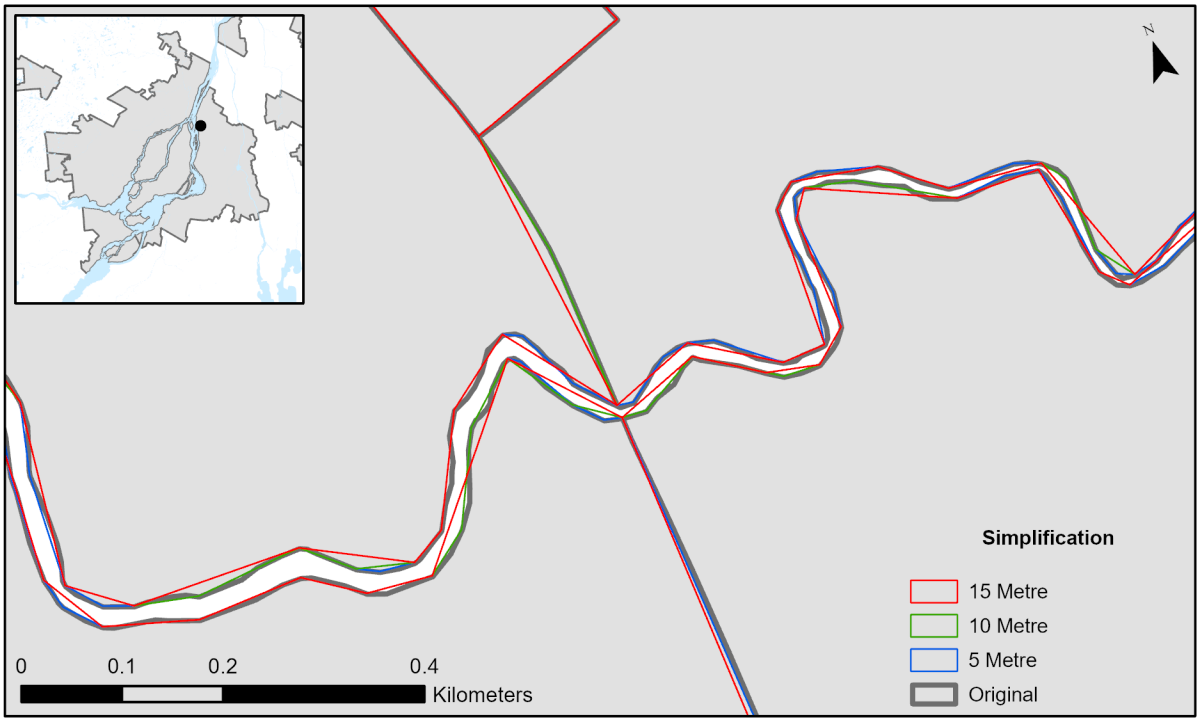
1. Topology correction

Several files for the 1971 to 1991 period had topology errors. These were corrected through different procedures depending on their prevalence: manually and using the “fill gaps” tool in ArcGIS Pro with a 1 km^2 threshold.

2. Polygon simplification

Created at different times using different tools, the complexity of the source files varies considerably. A significant proportion of this complexity occurs in ocean coastlines. To reduce file sizes and make polygons more comparable across years, we applied a topology-respecting simplification tool in ArcGIS Pro. The tool provides four simplification methods. The Retain Critical Points (Douglas-Peucker) method was selected because it substantially reduced the total number of vertices while retaining a high degree of fidelity to the original file. We also assessed several distance thresholds, ultimately using 10m. (**Figure 1** displays multiple thresholds in a Montreal-area example in 2021.) In some files, the number of vertices was reduced by as much as 80 per cent.

Figure 1: Assessing distance thresholds using Retain Critical Points simplification



3. Shoreline harmonization (CBF-Harmonized Shorelines series)

The harmonized files aim to make the handling of coastal shorelines and internal water bodies maximally consistent across all time. All files were clipped to a standardized shoreline. This was created by intersecting dissolved copies of the 1851 CSD file and the 2001 CSD file. The 1851 file was selected because they were consistent with the shorelines used across the 1851–1951 file series – the longest period of consistent shorelines across the full series. The standardized shoreline used to clip the files does not include any internal lakes or rivers. It does, however, include rivers that connect with coasts, including the Saint John River, the Fraser River, and the Ottawa River, as well as ocean-connected bays, such as Halifax's Bedford Basin and Hamilton Harbour, which are not removed from Statistics Canada's own CBFs. The intersection with the 2001 CSD CBF also removes many small islands in remote areas. Where DBF files are available, the CBF-Harmonized Shoreline series file was created from the DBF, not the CBF.

Next, the fill gaps tool was run to remove any remaining gaps in the polygon. This revised dataset was then simplified using the same technique described above. To remove any remaining geometry errors, a 0m buffer was created using the pairwise buffer tool with the geodesic or shape preserving method.

4. Attribute table standardization

The next step was to standardize the field names, order, and content of each file's attribute table (see **Table 2**.) Geometric attributes, including centroid and representative point coordinates and area were recalculated from the standardized geometries. The *geosid* field contains the geographic unit identifier contained in the source file.

Table 2: Standardized attribute table fields

Field	Description	Format
id	Polygon ID	numeric
geosid	Geographic "source" identification code from original file.	String
time	Census year	String "yyyy"
level	Level of geography (ct, csd, cd, cma, pr) Use these numeric codes: 1 = ct (nests within csd, cma) 2 = csd (nests within cd, pr) 3 = cd (nests within pr) 4 = cma (nests within nothing) 5 = pr (highest level)	integer
geoname	Geographic unit name - Use English name - If CT, use 0000.00 (ctuid without cmauid) if numbered; use string name if an unnumbered tract in 1951–66.	string
lat_c	Centroid coordinate – latitude (decimal degrees)	numeric
lon_c	Centroid coordinate – longitude (decimal degrees)	numeric
lat_r	Representative internal point coordinate – latitude (decimal degrees)	numeric
lon_r	Representative internal point coordinate – longitude (decimal degrees)	numeric
areakm	Area - square kilometres	numeric
source	Creator of original files (e.g., StatCan, CCRI, TCP, Allen, Hewitt etc.)	string
timestamp	Time exported (yyyymmdd)	String - time stamp
version	Version number (1, 2, 3, …)	numeric

5. Export files

The resulting files were then exported in multiple file formats and projections:

- Esri Shapefile in NAD83 CSRS / EPSG: 3348
- Esri Shapefile in WGS84 / EPSG:4326
- File Geodatabase in NAD83 CSRS / EPSG: 3348
- File Geodatabase in WGS84 / EPSG:4326
- Geojson in WGS84 / EPSG:4326

Project information: NAD83 CSRS / EPSG: 3348
North American Datum (NAD) 1983 Canadian Spatial Reference System (CSRS) Statistics Canada Lambert

- WKID: 3348 Authority: EPSG
- Projection: Lambert Conformal Conic
- False Easting: 6200000.0
- False Northing: 3000000.0
- Central Meridian: -91.86666666666666
- Standard Parallel 1: 49.0
- Standard Parallel 2: 77.0
- Latitude Of Origin: 63.390675
- Linear Unit: Meter (1.0)

Geographic Coordinate System (GCS): GCS North American 1983 CSRS

- Angular Unit: Degree (0.0174532925199433)
- Prime Meridian: Greenwich (0.0)
- Datum: North American 1983 CSRS
- Spheroid: Geodetic Reference System (GRS) 1980
- Semimajor Axis: 6378137.0
- Semiminor Axis: 6356752.314140356
- Inverse Flattening: 298.257222101