Catalogue no. 18-001-X
ISBN 978-0-660-75054-5

Reports on Special Business Projects

Mapping Location and Co-location of Industries at the Neighborhood Level: A Spatial Kernel Density Approach

by Jérôme Blanchet, Robert Oikle, Dennis Huynh and Alessandro Alasia

Release date: October 10, 2025

Statistics Canada Statistique Canada

Canada

How to obtain more information

For information about this product or the wide range of services and data available from Statistics Canada, visit our website, www.statcan.gc.ca.

You can also contact us by

Email at infostats@statcan.gc.ca

Telephone, from Monday to Friday, 8:30 a.m. to 4:30 p.m., at the following numbers:

- Statistical Information Service 1-800-263-1136
- National telecommunications device for the hearing impaired 1-800-363-7629
- Fax line 1-514-283-9350

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner. To this end, Statistics Canada has developed standards of service that its employees observe. To obtain a copy of these service standards, please contact Statistics Canada toll-free at 1-800-263-1136. The service standards are also published on www.statcan.gc.ca under "Contact us" > "Standards of service to the public."

Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued co-operation and goodwill.

Published by authority of the Minister responsible for Statistics Canada

An HTML version is also available.

Cette publication est aussi disponible en français.

Table of contents

Tables

Figures

Maps

Acknowledgement

The Data Exploration and Integration Lab (DEIL) and the Urban Data Lab (UDL) at the Center of Special Business Projects (CSBP) are grateful to Statistics Canada, particularly Christian Wolfe, Shujaat Ansari and Serge Godbout, for their knowledge of the Business Register, Dr. Mahamat Hamit-Haggar for coordinating the editorial process and internal reviews, Chris Li for the institutional review, Dr. Bjenk Ellefsen, for elaborating the vision and supporting the feasibility around the project, Dr. Ala'a Al-Habashna for initiating the data processing, and Zheng Yu and Wafa Ashraf for the final technical revision of the document. We are also grateful to Dr. Stephen Tapp and Patrick Gill from the Canadian Chamber of Commerce (CCC) for their review, comments, and suggestions.

User Information

Symbols

The following standard symbols are used in Statistics Canada publications:

.	not available for any reference period
..	not available for a specific reference period
...	not applicable
0	true zero or a value rounded to zero
0^s	value rounded to 0 (zero) where there is a meaningful distinction between true zero and the value that was rounded
^{p}	preliminary
^{r}	revised
x	suppressed to meet the confidentiality requirements of the Statistics Act
^{E}	use with caution
F	too unreliable to be published
*	significantly different from reference category ($p < 0.05$)

The following mathematical expressions are used in this paper:

$\mathrm{CMA}()$, Census Metropolitan Area IDs and main information

$\mathrm{DB}()$, Dissemination block IDs and main information

$\mathrm{DA}()$, Dissemination Area IDs and main information

KDE, Kernel density estimation

$\mathrm{PKDF}()$, Polynomial kernel density function

$\mathrm{G}()$, Grid of output cells or tiles

$\mathrm{G}\left(\mathrm{mDB}\right)$, Grid of median DBs

$\Phi()$, Grid output cell IDs and main information

$\Phi\mathrm{c}()$, Grid output cell centroid

$\psi\ ()$, Kernel density bandwidth

$\mathrm{m}()$, Minimum of some items

$\mathrm{X}()$, Sample distribution

$\mathrm{s}()$, Sample size

$\sigma()$, Standard spatial distance from mean center

$\mu()$, Median spatial distance from mean center

$\mathrm{IQR}()$, Interquartile range 75%-25%

$\mathrm{L1}()$, Latitude numerical coordinate

$L2()$, Longitude numerical coordinate

$L1C()$, Latitude numerical coordinate of mean center of CMA

$L2C()$, Longitude numerical coordinate of mean center of CMA

$\sqrt{A}()$, Side length of a median DB

$\pi\psi^2()$, Total spatial superficies covered by bandwidth around grid output cell centroid of interest

i, Index of items of included in spatial superficies covered by bandwidth around grid output cell centroid

$n()$, Total number of items of X included in spatial superficies covered by bandwidth around output cell centroid

$JW()$, Job unity weight

$JL()$, Job random spatial location

$\Theta()$, Distance between random job location and grid output cell centroid of interest

$f = \frac{3}{\pi} r()$, Main ratio of polynomial kernel density function (individual density contribution)

$\sim U(DBP)$, Uniform spatial distribution over a dissemination block polygon

i.i.d, Independent and identically distributed

RP, Multinomial random process for the methodology

$E(RP)$, Expected event for multinomial random process of the methodology

$T(RP)$, Tail rare event for multinomial random process of the methodology

BRP, Binomial random process for the analogical example

$E(BRP)$, Expected event for the binomial random process of the analogical example

$T(BRP)$, Tail rare event for binomial random process of the analogical example

b, Large number of small dissemination block for binomial random process of the analogical example

v, Large and even number of jobs for binomial random process of the analogical example

CLT, Univariate central limit theorem

$(->N())$, Convergence toward normal distribution

$\sum_{j=1}^{v}$, Summation of a finite number of v items

$C(v,j)$, Number of ways of combining j item among v item without order

!, Factorial

$J()$, Set of dimension n x 2 for $JW()$ and $JL()$ over $n()$

y, Total density value from the finite sum of individual density contribution

x, Ratio of distance to bandwidth $\Theta()/\psi()$

df/dx or f', First derivative of

d^2f/dx^2 or f'', Second derivative of f

$\in(,),\in[,]$, Item included in open and bounded interval

$-\infty,+\infty$, Minus and positive infinite

MCLT, Multivariate central limit theorem

S, Large finite number of spatial spot available on a DB

p, S-dimension vector of S probabilities summing to 1

vp, vector of S entries representing the expected number of jobs allocated to each of the S spatial spot of the DB

vM, finite covariance matrix of dimension S x S of multivariate normal distribution of our converging uniform multinomial random process

$\left(\mathrm{vM}\right)^{-1}$, Inverse of matrix vM

I, Identity matrix of dimension $\mathrm{S\ x\ S}$

P, Diagonal matrix whose diagonal elements are the items of vector p

$\mathrm{SVN}\left(\mathrm{x'};\mathrm{vp},\mathrm{vM}\right)$, Multivariate or S-variate normal distribution

$\det\left(\mathrm{vM}\right)$ or $\left|\mathrm{vM}\right|$, Determinant of matrix vM

k, Factor of density reduction of uniform spatial points on a plane

$\mathrm{g}\left(\mathrm{k}\right)=\mathrm{k}^{1/2}$, Average distance expansion of uniform spatial points on a plane

Δ, Small positive variation of factor k

$\approx$, approximation

Mapping Location and Co-location of Industries at the Neighborhood Level: A Spatial Kernel Density Approach

by **Jérôme Blanchet, Robert Oikle, Dennis Huynh** and **Alessandro Alasia**

Executive summary

Businesses don't just choose a city to locate their operations; they choose a neighbourhood[1]. In turn, the clustering of businesses in a neighbourhood shapes the economic opportunities and quality of life in the area. Despite the relevance of these local-level dynamics documented in the literature, research on business clusters in Canada has focused primarily on the regional or metropolitan scale. This focus has limited the possible applications of cluster analysis for urban planning, infrastructure development, and local development by actors operating or delivering programs at the local scale.

The steady improvement in the geolocations of business data is providing new analytical opportunities. This paper presents a method to define business clusters at a granular sub-metropolitan level. Using data from Statistics Canada's Business Register (BR) for selected industries, employment locations at the establishment level are spatially distributed within their respective dissemination block (DB) (a block in urban and rural areas). A spatial kernel density estimation (KDE) approach is performed on these employment locations to define the boundaries of business clusters. A new approach to define the kernel bandwidth is detailed since the traditional Silverman's rule bandwidth method fails, in the case of our applications, to directly recognize the configuration of the DB structure within the cities of interest. Results are developed for three industries (manufacturing, retail trade, and accommodations and food services), as well as some industry clusters as defined by Delgado et al. (2014), for four major metropolitan areas (Montreal, Toronto, Winnipeg, Vancouver).

The results are mapped for each type of cluster and metropolitan area showing different spatial configurations for different industry sectors. As expected, retail trade and accommodation and food services clusters are relatively more scattered across metropolitan areas, compared to manufacturing clusters. However, simple statistics on establishments and employment counts show that the geographic boundaries of neighborhood clusters generated by this analysis capture most of the employment and establishment counts located in the Census Metropolitan Area (CMA) of reference and associated with the industry cluster. For instance, the manufacturing sector cluster contains 89.7% (Montreal), 94.4% (Toronto), 91.9% (Winnipeg), and 90.8% (Vancouver) of total manufacturing employment in the respective CMAs. The results also point to greater co-location of specific types of business, such as retail trade and accommodation and food services. This preliminary analysis appears promising in revealing patterns of business co-location in defined neighborhood areas.

The methods used to define these neighbourhood-level clusters open new opportunities for timely analysis of business conditions at the local level, as well as broader types of analyses at the neighborhood-level (e.g., social disparities, and quality of life), accounting for the business composition of the area. The use of geographic boundary files of specific business clusters, as a geofencing tool, can be employed to monitor local business performance and trends in combination with other Statistics Canada data holdings or alternative data sources, such as mobility flows.

This project proposes an experimental methodology and a set of experimental industry clusters. People's feedback is valued. Commentaries, and suggestions can be communicated to the lead author, Jérôme Blanchet (819-576-5502), Unit Head at the Data Exploration, and Integration Laboratory (DEIL) and Urban Data Laboratory (UDL), of the Centre for Special Business Projects (CSBP), Statistics Canada.

1. See: Bridges, Winter 2005-2006 (Federal Reserve Bank of St. Louis)

Introduction

The academic and policy debate on business clusters has spanned over three decades. In most of the ensuing literature and policy documents, business clusters are defined as a geographic concentration of interconnected enterprises, organizations, and institutions within a specific industry or sector (Wolfe and Gertler, 2004). Theory and evidence suggest that spatial proximity and agglomeration facilitate collaboration, linkages, resource sharing, and other synergies. Thus, policy support for clusters draws on the idea that the proximity of businesses and supporting organizations in a particular geographic area fosters synergies, innovation, and competitive advantages (Bekar and Lipsey, 2001).

The cluster literature in Canada has delved into the effect of clusters on business performance, employment, and wages (Lucas et al., 2009; Niosi and Bas, 2001; Steiner and Ali, 2011; Spencer et al., 2010). It has examined policies in support of clusters, assessed their effectiveness (Niosi and Bas, 2001), and it has explored methodologies to identify and map clusters across space (Spencer, 2014). Overall, the insights generated by this literature are that spatial clustering of businesses creates an environment conducive to innovation, resource efficiency, collaboration, and overall economic development. These clusters contribute to the growth and success of individual businesses while enhancing the competitiveness and resilience of regions and metropolitan areas.

Most of the Canadian literature has focused on a regional scale or on individual metropolitan areas, and for many applications and policy purposes, an analysis at the regional level (city, metropolitan area, or labor market area) will remain adequate. Nevertheless, there is a growing number of applications which require neighborhood-level analysis and provide unique insights to both local-level actors (municipalities and other local business organizations) as well as to provincial or federal actors. The demand for geographically granular data on business conditions and trends is constantly growing from federal agencies and local stakeholders, such as municipalities, business organizations, and the business community.

This paper brings the analysis of business clusters to a more granular geographic scale by developing a methodology for identifying business clusters at the neighborhood level. The proposed method identifies clusters of businesses at the DB level, which is one of the most granular spatial units of analysis defined by Statistics Canada.[2] The method is developed with an application to four census metropolitan areas (CMAs) of different sizes and for different industry cluster specifications, including simple 2-digit North American Industry Classification System (NAICS) groups as well as industry clusters resulting from groupings of NAICS codes, as defined by Delgado et al. (2014).

The increasing accuracy of geolocation for establishments in the BR of Statistics Canada enables this analysis, while the possibility offered by business microdata linkages creates opportunities to explore a multitude of business performance dimensions.[3] In this context, one of the main challenges in this type of analysis is to preserve the confidentiality of sensitive business information while providing valuable insights to the business community and policymakers. Hence, this paper remains a first exploration at the intersection of the highest level of granularity and confidentiality preservation of business information.

This paper is organized into five main sections. The next section presents a cursory review of the existing literature on cluster analysis in Canada, intended to highlight the information gap at the neighborhood level and the main motivational aspects for this research, as well as examples of neighborhood-level analysis from other countries. This is followed by a detailed presentation of the data and methodology applied in this analysis, which involves a new approach for the kernel bandwidth calculation. The next section presents selected results and validates the findings, followed by a discussion on further development and possible applications of the cluster delineations. Finally, the annex includes a large supplementary content of 23 high-resolution cluster maps.

Why a neighborhood dimension of clusters

Most studies on business clusters in Canada use regions or cities as geographic units of analysis. For many applications and policy analyses, that geographic scale provides an adequate level of granularity. Business performance indicators have become relatively abundant at the municipal or regional scale. Within cities

2. See: Dictionary, Census of Population, 2021 – Dissemination block (DB) (statcan.gc.ca)
3. See for instance, the possibilities offered by the Business Linkable File Environment, at: Business - Linkable File Environment (statcan.gc.ca)

or regions, it is assumed that business proximity is sufficient to allow for the interactions that underpin the very concept of clusters and the benefits that derive from it. Some of the well-known clusters, for instance, Silicon Valley in Northern California, and the Research Triangle Park in North Carolina, are spread over several municipalities and a regional scale appears appropriate to study the dynamics and development of these clusters.

Nevertheless, there is evidence of clusters, such as financial, cultural, retail, or manufacturing clusters, concentrating in specific neighborhoods within a metropolitan area. Similarly, there is evidence that spatial disparities and differences in business performance may be as pronounced at neighborhood scale as they are at the regional scale (Wheeler, 2006; OECD, 2018). Several stakeholders and policymakers, such as local business organizations and municipalities, are operating with a neighborhood lens and are developing or delivering policies that impact businesses in specific areas within a municipality. Hence, these stakeholders require a more geographically granular analysis of clusters and cluster performance.

To respond to these information needs, different streams of research have analyzed business clusters at the neighborhood level, looking at location choices within metropolitan areas, the impact of municipal regulations and policies on the formation and growth of clusters, the role of local associations, and the impact and economic spillover of clusters on the surrounding neighborhoods. This research comes from different disciplinary perspectives. Ranging from the more traditional analysis of business clusters to urban planning studies, and applied research and analysis generated by local associations, boards of trade, and municipal planning departments. The remainder of this section provides an overview of this literature, highlighting key insights and pointing to the current information gap in the Canadian context.

Spatial disparities within cities, at neighborhood level, are a well-known and researched phenomenon (OECD, 2018). These disparities are reflected and exacerbated by business location choices and clustering in different neighborhoods. For example, Wheeler (2006) shows that positive business growth in the St. Louis metropolitan area, in the US, is, in fact, the result of substantial business growth in one neighborhood combined with a decline in another neighborhood. These dynamics are driven by both neighborhood and business characteristics. These insights are relevant from both a business perspective (for location choices), as well as from a municipal perspective (for policy to support business clustering and reduce neighborhood socio-economic disparities). The importance of a neighborhood dimension of spatial clusters is also highlighted by Gabaix (2011), who uses satellite imagery data to define clusters of population density. These results suggest that data generated from cluster density-based approaches for city boundaries are easier to integrate within a spatial model compared to regions defined by administrative boundaries directly.

Insights from analysis on business clusters at the neighborhood level are related to the location choices of businesses. Wheeler (2006) notes that prospective investors do not choose just a region or city, they also choose a neighborhood, which in some contexts may refer to a specific industrial park or business district. Therefore, understanding the dynamics of industrial parks[4] within a metropolitan area, and their surroundings, is of specific relevance. Similar insights come from Arauzo-Carod (2021), who examines the location choices of high-tech firms at the neighbourhood level in Barcelona. This analysis shows that both neighbourhood characteristics and amenities matter in business location choices and that spatial spillovers are relevant for some high-tech industries.

Business clusters do not just determine economic composition and job availability in a neighborhood. Several studies highlighted that they may also shape the quality of life of neighborhoods and their attractiveness for residential use (Shybalkina, 2022; Stern and Seifert, 2010). Research has studied the impact of specific clusters on neighborhoods, particularly arts and culture clusters within metropolitan areas. Grodach et al. (2014) identify arts clusters at the regional and neighborhood levels, using Zip Codes, for US metropolitan areas of different sizes. Moreover, their findings show that arts industries exhibit distinct metropolitan area and neighborhood-level location patterns. It is also revealed that while many of the characteristics of arts clusters are place specific, the arts are associated with broad measures of local innovation and development, suggesting that these business clusters can play a larger role in economic development for metropolitan areas.

Another stream of literature on intra-urban business clusters focuses on the delineation and role of business districts. This literature has often been categorized under the heading of central business districts (CBD) analysis (Meltzer, 2012; Yu et al., 2015); it has both methodological and policy relevance, with part of this literature focusing

4. Industrial park - Wikipedia

on the modeling aspects, while other parts delving into the role and dynamics of organizations related to the management, development, and promotion of these business districts. The methods used in delineating CBD range from use of remote sensing data (Taubenböck et al., 2013) to census data on job density (Yu et al., 2015).

The approach taken in the present analysis is inspired by this literature and, in particular, by the analysis of Sergerie et al. (2021), which aimed to develop a method to identify the geographic boundaries of Canada's downtown neighbourhoods. Sergerie et al. (2021) use spatial KDE to calculate a density surface of job location data at the dissemination area (DA) level, allowing for comparisons of these areas across Canada.

Business cluster formation, at the neighborhood level, like at the regional level, is not just a spontaneous process or a cumulative result of historical accidents. That is, it is not a random phenomenon, and it can be explained. Municipalities play a key role in shaping, developing, and supporting the clustering of businesses in specific areas (Zhang, 2019). Typically, this is done through zoning, which is the regulatory method used by municipalities or local governments to set rules that define the activities and buildings that can occur in a certain location. In this way, municipalities provide the space, infrastructure, and services for the development of industrial parks, as areas dedicated to industrial uses.

Like municipalities, local business associations are key actors in supporting the development of local business clusters (Dhamo et al., 2023). In Toronto, for instance, the Toronto Board of Trade (2021) released a study mapping five types of district areas across the broader metropolitan region, which include a metro center, goods production and distribution areas, services and mixed-use areas, regional centers, and knowledge creation centers. In parallel, that municipality is home to 84 Business Improvement Areas (BIAs). These are local associations of businesses aiming to support competitive and attractive business areas for consumers and new businesses.

The proactive role that local actors can take in shaping economic development and prosperity of their area explains why neighborhood-level analysis is becoming increasingly relevant. Analyzing local clusters or concentrations of businesses within a neighborhood can provide insights into consumer behavior, local employment, and the overall economic health of the neighborhood. These local actors operate in geographically defined ecosystems for which data may be generated from local sources or analyzed at the local level. What is missing, particularly in the Canadian context, is the broader and comparative framework that would allow neighborhood cluster analysis nationwide, with standardized definitions across jurisdictions, which is the information gap that this paper is intended to fill. Improvements in georeferencing of business microdata and advancements in spatial data analytics with large databases are making it increasingly feasible to conduct fine-grained analyses at the neighborhood scale.

A proposed methodology

The general approach used in this analysis is drawn from the work of Sergerie et al. (2021) on the definition of downtown areas for metropolitan areas in Canada. That analysis applies spatial KDEs on the geolocation of total jobs derived from the place-of-work status variable of the Census of Population; the geographic unit of analysis in that application is the DA.

In the present study, business clusters are defined with an analogous method, using spatial KDEs applied to the geolocation of employment recorded at the establishment level. Data on establishments are extracted from the BR for selected industries, and the unit of analysis is the DB, a more granular unit than the DA. Given the substantially greater level of granularity (both for selected industries and geography), the methodology used to define business clusters presents several additional steps compared to the workflow outlined by Sergerie et al. (2021). The following sections describe the proposed methodology in detail.

Study Areas

Four study areas were selected for the development of the methodology, representing different levels of urban density in Canada. These study areas are the CMA of Montréal, Toronto, Winnipeg, and Vancouver. Each CMA comprises of a different number of municipalities (census subdivisions), with neighborhoods reporting substantially different population and employment densities and degrees of urbanization.

The geographic unit of analysis used for the geolocation of businesses is the DB. A DB is an area bounded on all sides by roads and/or boundaries; that is, in urban areas, it is what is commonly referred to as a block. DBs are part of Statistics Canada's standard geographic areas for dissemination, and they are the smallest geographic area for which population and dwelling counts are disseminated with coverage for all the territory of Canada.[5]

Business Register

The data used in the analysis come from Statistics Canada's BR, which is the continuously maintained central repository of baseline information on businesses and institutions operating in Canada. For this analysis, data are for the reference period of December 2023.

In the BR, industry sectors are defined by NAICS codes. The use of BR data presents advantages when compared to other possible data sources, such as place-of-work data from the Census of Population. Establishment-level data from the BR is classified with more detailed NAICS (6-digit codes), which allows for detailed custom clusters. Employment data in the BR are also updated with greater frequency. Although they are not as accurate as dedicated employment statistics, they are a timelier alternative to census data.

The NAICS codes included in this analysis represent six different industry clusters. Three industry clusters are composed of 2-digit NAICS codes, and the remaining are defined according to the work by Delgado et al. (2014). These industry clusters and their corresponding NAICS codes are summarized in Table 1.

The industry clusters generated at the 2-digit NAICS level, include the Manufacturing Sector (NAICS codes 31, 32, and 33), the Retail Trade Sector (NAICS codes 44 and 45), and the Accommodations and Food Services Sector (NAICS code 72).

The three industry clusters defined according to Delgado et al. (2014), are as follows. First, the Distribution and Electronic Commerce (cluster 10); this cluster consists primarily of traditional wholesalers as well as mail-order houses[6] and electronic merchants. The companies in this cluster mostly buy, hold, and distribute a wide range of products such as apparel, food, chemicals, gases, minerals, farm materials, machinery, and other merchandise. The cluster also contains firms that support distribution and electronic commerce operations, including packaging, labelling, and equipment rental and leasing. The second cluster is Financial Services (cluster 16); this cluster contains establishments involved in aiding the transaction and growth of financial assets for businesses and individuals. These firms include securities brokers, dealers, and exchanges; credit institutions; and financial investment support. Insurance firms are located in a separate Insurance Services cluster. Finally, the Hospitality and Tourism (Cluster 22) contains establishments related to hospitality and tourism services and venues. This includes sports venues, casinos, museums, and other attractions. It also includes hotels and other accommodations, transportation, and services related to recreational travel, such as reservation services and tour operators.

The hierarchical structure between the 3 industry clusters generated at the 2-digit NAICS level, and the 3-industry cluster defined according to Delgado et al. (2014) is not straightforward. That is, the several 4-digit NAICS level used according to Delgado et al. (2014) are not necessarily composed of the first 2-digit 31, 32, 33, 44, 45 and 72. Furthermore, for space convenience and simplicity, the recurrence of the 6 industry clusters is not necessarily consistent across this paper. That is, some results analysis and methodology steps focus on the 3 industry clusters generated at the 2-digit NAICS level alone, while some other focuses on the 6 industry clusters.

5. See: Illustrated Glossary - Dissemination block (DB) (statcan.gc.ca)
6. Mail Order Houses: Meaning, Advantages, and Disadvantages | GeeksforGeeks

Table 1
Industry clusters with associated NAICS codes

Industry cluster	NAICS codes included
Manufacturing Sector	31, 32, 33
Retail Trade Sector	44, 45
Accommodations and Food Services Sector	72
Distribution and Electronic Commerce (cluster 10)	4111, 4121, 4131, 4132, 4133, 4141, 4142, 4143, 4144, 4145, 4161, 4162, 4163, 4171, 4172, 4173, 4179, 4182, 4183, 4184, 4189, 4191, 4232, 4234, 4235, 4236, 4238, 4239, 4241, 4242, 4243, 4244, 4245, 4246, 4247, 4248, 4249, 4251, 4541, 4931, 5324, 5614, 5619
Financial Services (cluster 16)	5211, 5221, 5222, 5223, 5231, 5232, 5239, 5259, 5269, 5614
Hospitality and Tourism (cluster 22)	1142, 4539, 4871, 4872, 4879, 5322, 5615, 7112, 7121, 7131, 7132, 7139, 7211, 7212, 7213

Source: authors' computations from the BR database.

Density Estimation

The spatial concentration of industries was determined using a spatial KDE method. The spatial KDE method is a non-parametric technique that estimates the probability density function of a random variable over a spatial domain. For each metropolitan area, clusters were identified using KDE results by aggregating adjacent DBs with a minimum density of employment in each industry or combination of industries.

Employment counts from establishment-level data of the BR were geocoded to the DB spatial boundaries files. The total number of employees in each DB was then calculated. Job locations representing each employee were randomly and uniformly distributed within the DB boundary[7]. This processing step must be acknowledged, not as a lack of accuracy and weakness of the data, but as a way to leverage the methodology forward. That is, this approach smooths the relatively sparse spatial distribution of jobs within the DB, and eventually facilitates the KDE process by generating a relatively more continuous distribution.[8] The uniformity argument was made to not prioritize any sub-region of the DB during the randomization process. That is, the concentration of establishment density within some specific spots of the DB doesn't affect the job randomization process. Some sections of this report below describe in more detail the randomization process.

The density estimation section of this paper is made of two sub-sections. First, we document partial information about the Polynomial Kernel Density Function, with a focus on the necessary arguments to understand the use and significance of the kernel bandwidth. We also explain why the Silverman kernel bandwidth approach is not a good fit for our application, and then, we describe a new method for the computation of the bandwidth. Second, we describe the full information about the Polynomial Kernel Density Function. That is, all remaining arguments of the function not described so far. We also document the random processes used.

Kernel Density Bandwidth Methodology and Other Parameters

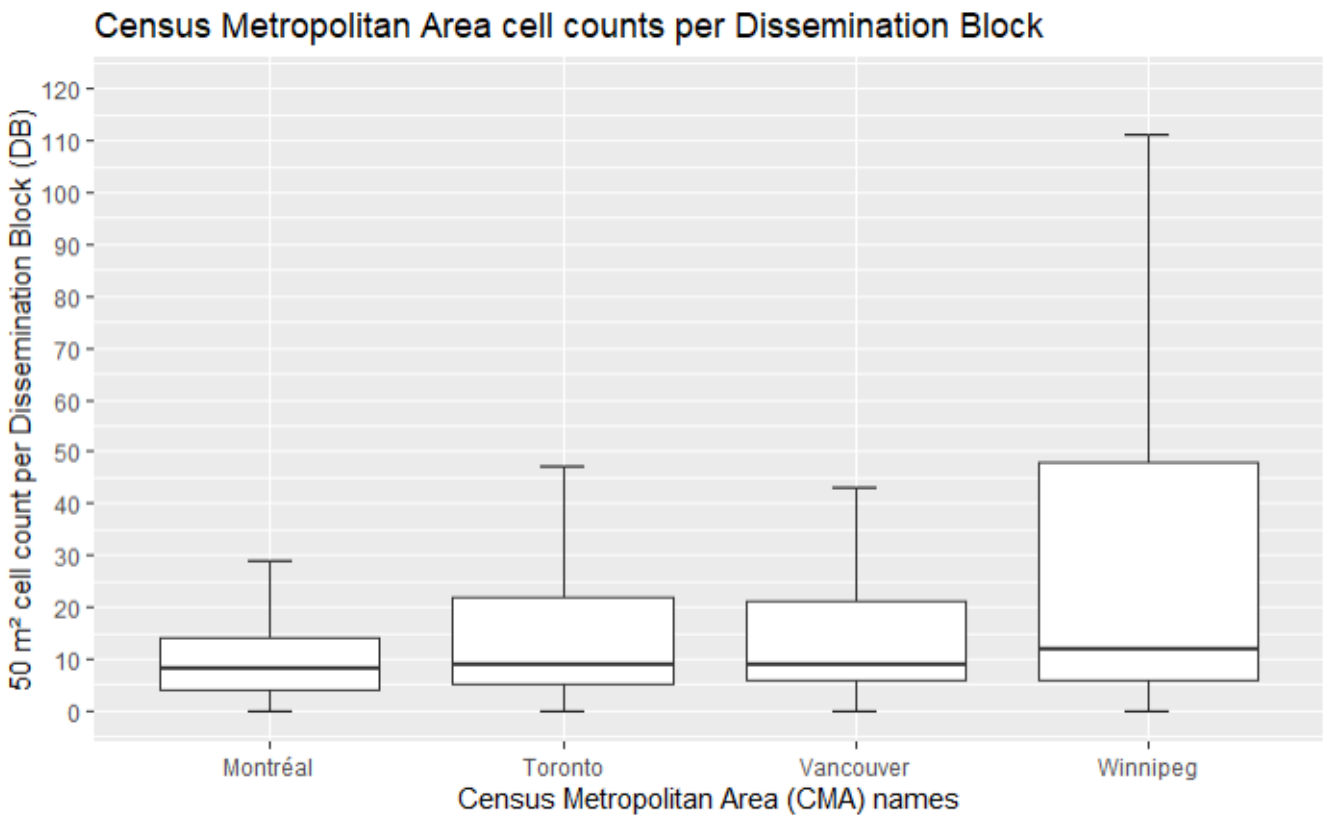
Before proceeding to the full KDE description, we specify partial details of the Polynomial Kernel Density Function, $\mathrm{PKDF}(\Phi,\Phi c,\psi)$, which is the main instrument of the KDE approach. The function involves a minimum of 3 arguments: the grid output cell id, Φ, the grid output cell centroid, Φc, and finally, the bandwidth or radius, . The polynomial functional form was selected according to Sergerie et al. (2021), even though a normal and uniform form was also available. Φc was defined geometrically, and not weighted according to the establishment spatial density concentration of the DB. Sets of specifications and testing for, Φ and ψ were considered for economic interpretability of the results and computational efficiency matter. Those are briefly documented below.

The spatial resolution, or dimension of a cell (or tile) for a grid G is a parameter of choice. For the purpose of this analysis, a squared grid with a cell dimension of 50x50 metres was adopted. Following the testing of different specifications (e.g., a range of 10x10 metres to 100x100 metres), the 50x50 metres grid was adopted to strike a balance between spatial detail and computational efficiency. Figure 1 below shows the distribution of the number

7. Annex 2 of this paper propose an alternative randomization method where the jobs of a DB are randomly distributed within the DB, DA and Aggregated Dissemination Area (ADA).
8. The limit of this approach arises for large DBs that contain a single establishment or few establishments with many employees located in a restricted part of the DB. In such cases, this spatial configuration would be superseded by the random distribution of jobs within the DB. Overall, however, the use of DBs instead of DAs mitigates these challenges, which are typically posed by administrative geographies. The DB is a more granular administrative geography than the DA and provides a closer approximation to the neighborhood concept.

of 50x50 metres cell per DB, in the four metropolitan areas. The Figure confirm that a minimum and reasonable number of cells are available in most of the DB of each CMA. Furthermore, the median number of cells per DB ranges from 7.5 to 12.5. Finally, the Figure also shows variations across CMAs for the range of total number of cells that can fit within a single DB. This variation is particularly important, since it captures the characteristics of both large and small CMAs within Canada. That is, smaller CMAs generally involves a larger range of DB dimension, while the largest CMAs of our study are made of a smaller and lower range of DB dimension. It is worth noting that Montreal is the only CMA to get a symmetric distribution of DBs, as the opposite of the 3 other CMAs where the distribution is relatively more skewed.

Figure 1
Total number of output cell per DB for the four study areas



Source: authors' computations from the BR database.

Similarly to the choice of grid cell dimension, the choice of the kernel bandwidth, ψ , also has implications for the results. To understand the relevance of the bandwidth, consider the analogy to a simple histogram. Excessively large bandwidths, i.e., a histogram with a few numbers of bars, will mask the underlying distribution. On the contrary, a bandwidth that is too small may result in a frequency of one unit for every outcome of the distribution, hindering a clear understanding of the data distribution with a flat histogram.

To define ψ , various specifications were tested, including the well-known B. W. Silverman's Rule of Thumb's Silverman, (1986) as applied in Sergerie et al. (2021). Given the grid cell size and DB configuration used in our analysis, the Silverman's bandwidth did not succeed to include enough data content in its local surroundings. Consequently, the Silverman's process was not effective in smoothing the spatial discrete distribution of job locations, making the KDE ineffective and resulting in fragmented clusters. The following paragraph proposes an interpretation about the reasons for the Silverman's process to fails in providing a satisfactory bandwidth for the case of our applications in this research.

The Silverman equation captures the dispersion of the job locations around a point of reference within the CMA. From the previous section, we know that the job locations are originally geo-coded within their respective establishment fixed spatial locations and are then distributed randomly and uniformly within the boundaries of their respective DB, without prioritizing the location of the establishment or any sub-region of the DB. Therefore, we can state that the Silverman equation only encapsulates partial information about the distribution of DB dimension within the CMA. However, the equation doesn't necessarily acknowledge the dimension of a typical DB, that is, an average or median DB. Consequently, in the specific examples of our applications, the Silverman bandwidth fails to aggregate the data across DBs and only propose kernel data transformation within the boundaries of the DB of reference, leaving the final DB level results unchanged compared to the original BR data.

We quickly provide an interpretation for this problem of why the bandwidth is too small. The Silverman bandwidth $= 0.9\mathrm{m}/s^{1/5}$, where s is the sample size and $\mathrm{m} = \min\left(\sigma, \left(1/\ln\left(2\right)\right)^{1/2} * \mu\right)^{1/2}$, where σ is the standard dispersion of the spatial job data points distribution X to the unique mean center point of the CMA, and where μ is the median dispersion statistical moment (50th percentile) in the distribution of dispersion of X to the mean center. Note, another version of the Silverman equation substitutes the median μ for the interquartile range $\mathrm{IQR}()$, where $\mathrm{IQR}() = 75\%$ percentile minus the 25% percentile.

Formerly, the unweighted equation for the standard distance is,

$$\sigma = \sqrt{\left(\sum_{j=1}^{s} \frac{\left(\mathrm{L1}(j) - \mathrm{L1C}\right)^2}{s} + \sum_{j=1}^{s} \frac{\left(\mathrm{L2}(j) - \mathrm{L2C}\right)^2}{s}\right)}$$

where $\mathrm{L1}$ and $\mathrm{L2}$ are the longitude and latitude numerical coordinates of the elements of X, respectively, and where $\mathrm{L1C}$ and $\mathrm{L2C}$ are the longitude and latitude numerical coordinates of the mean CMA center, respectively. Intuitively, both σ and μ cannot be contained within a typical or median DB of the CMA because the job locations are largely available spatially across the whole superficies of the CMA. Consequently, we assume the minimum (min) of the 2 quantities to be reasonably large enough and not able to explain a too small Silverman bandwidth. On the other hand, the sample size s is at the denominator of the Silverman formula and s can provide a very small bandwidth if s is too large. BR distribution of employees' number per establishment can be extremely right skewed, including positive outliers reaching very large values and adding up to a very large s due to the potential heaviness of the right tail. Consequently, if s is large enough to keep $s^{1/5}$ large enough, then the bandwidth will be unrealistically small. This research doesn't dedicate time to an outlier removal process for the BR data, but could be the objective of a future study. Specifically, the application of papers published by Statistics Canada, such as, Outliers in Sample Surveys by Lee et al. (1992), A Cautionary Note on Clark Winsorization by Mulry et al. (2016), A Method of Determining the Winsorization Threshold, with an Application to Domain Estimation by Martinoz et al. (2015), and On Searls' Winsorized Mean for Skewed Populations by Rivest et al. (1995). It is also worth noting that the Silverman equation assumes normally distributed kernel density values, which is not the case for our highly right skewed density distribution presented above. Finally, we point out that Mathematica uses a modification of the Silverman's equation, that is, $0.09 * \sigma$, for large sample size s, reaching beyond 100,000 observations. However, we did not use analytical software Mathematica for this research and decided to not leverage this version of the Silverman equation. Intuitively, taking off the large sample size s of the BR at the denominator of the Silverman ratio would contribute to resolve the issue of a too large denominator and a too small bandwidth, if substituting a coefficient of 0.9 with 0.09 (around 90% reduction) is not reducing the bandwidth too much.

To address the problem related to the Silverman's rule of Thumb in our applications, this research develops its own bandwidth methodology that focuses on the distribution of DB superficies provided by the data. In other words, our custom bandwidth accounts for the dimension of the median DB in the CMA of reference.[9] This new approach ensures that the bandwidth would be sufficient to span across both a typical DB, and the neighbouring DBs. This condition is fundamental, because the final product of this project is at the DB level. Consequently, if density for all grid cells represents the aggregation of information located within the boundaries of their respective DB only, there would be no exclusive spatial information generated from this research.

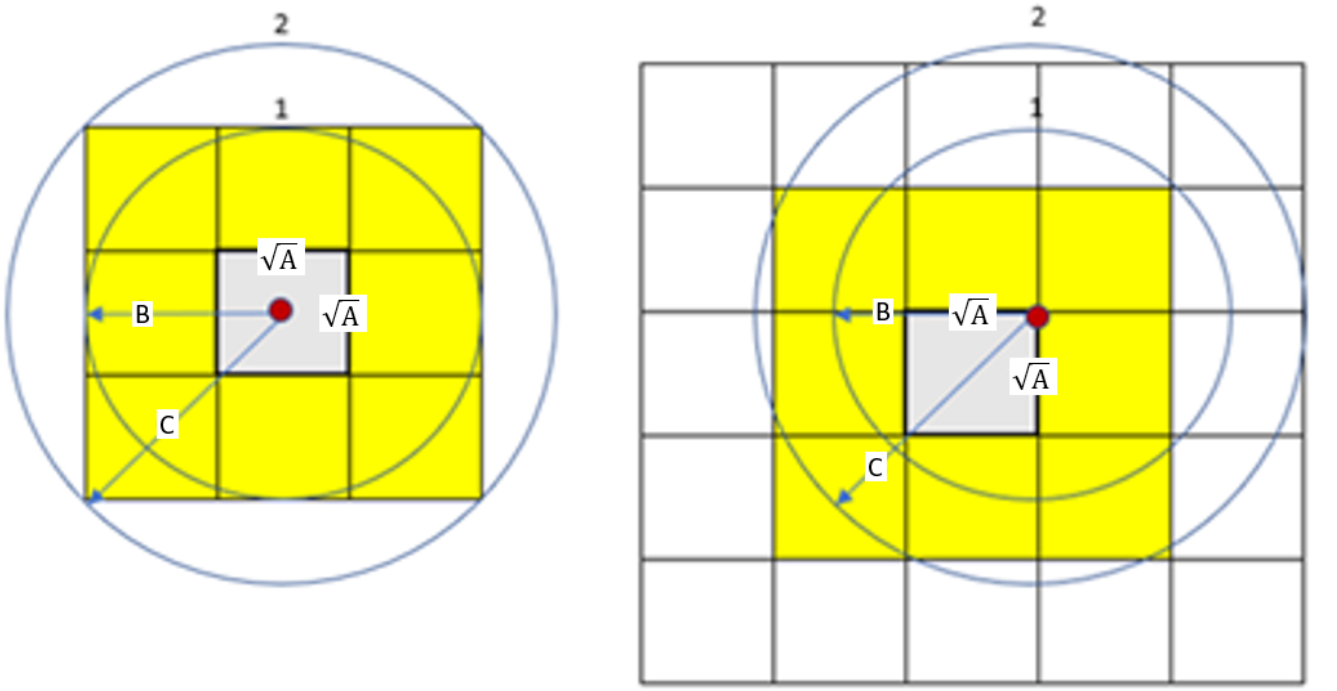
Figure 2 illustrates the bandwidth approach developed for the present analysis. The grid $G(mDB)$ populated by DBs, not to be confused with the grid of output cells G introduced above, assumes a local environment with equal squared DBs whose dimensions are those of the empirical median DB of the reference CMA. The median DB is derived from a projected area in the coordinate reference system (CRS) EPSG:3347. $\sqrt{A}$ is the side length of a median DB. Consequently, $\sqrt{A} * \sqrt{A}$ is the volume or superficies of a median DB. The red points are the geometric centroids of grid cells within a median DB of $G(mDB)$, and should not be mistaken as the DB's centroid.

Based on the left-sided example of Figure 2, bandwidth #2, represented as a circle, succeeds in reaching out to the totality of the requested information, which is highlighted in grey and yellow. That is, the full reference DB of the cell and all 8 direct neighborhood DBs of the reference DB, respectively. It is worth noting that it is the smallest possible bandwidth to meet these conditions. Specifically, it covers $(14.137 * A) / (9 * A) = 1.57$ times the requested superficies. The calculation of the bandwidth length uses a simple Euclidean distance logic; $B = \sqrt{A} * 1.5$, therefore, $C = \psi = \sqrt{(B^2 + B^2)}$. This logic makes sure that a grid output cell centroid overlapping at the DB's geometric centroid will be able to capture all needed data. Furthermore, based on the right-sided example of Figure 2, the grid output cell is now located at the top-right corner of the same reference DB. Bandwidth #2 succeed to reach out to all request DBs, except for 1 DB, at the bottom left, whose superficies is reachable at 50%. This unreachable area is acknowledged as acceptable in our methodology considering new DBs now become reachable at the top-right, even though they are not a direct neighborhood of the reference DB.

Cluster heat maps produced using this new bandwidth approach yielded a smooth surface while preserving neighborhood details (see details in the annex). For clarity, this paper acknowledges the bandwidth notation as $\psi(MDB(CMA))$, since the custom radius is now dependent on a single argument, the Median DB (MBD) configuration, and the MDB is itself specific to the CMA. That is, the bandwidth changes across CMAs but it is fixed within CMAs, no matter the variance of concentration of job locations in the neighborhoods of the CMA. However, notation ψ will remain in use for space efficiency matter. Finally, it is worth noting that our custom bandwidth, ψ, is not dependent on the dispersion of the job locations. Consequently, if CMA A and B have the same MDB and CMA B has twice the dispersion of job locations than CMA A, then both CMAs will get the same ψ, which is improper because the dispersion of job location matters. On the other hand, our custom bandwidth is robust to large outliers in the distribution of BR's number of employees per establishment. An improvement of our bandwidth would include both a dependence on the MDB and job location dispersion.

9. Since the distribution of DB dimensions are typically skewed, the median DB of the distribution was used instead of average DB, to be more robust to outliers.

Figure 2
Kernel Density bandwidth calculation for the understanding of the DB dimension distribution



Legend: $\text{Side Lenght of a Median DB} = \sqrt{A}$, $A = \text{Median DB Area} = \sqrt{A} * \sqrt{A}$, $B = \sqrt{A} * 1.5$, and $C = \psi = \sqrt{\left(B^2 + B^2\right)}$

Source: authors' design and methodology.

Uniform Multinomial Random Processes and Polynomial Kernel Density Function

Following the specification of arguments Φ, Φc and Ψ , this section documents the remaining details of the spatial polynomial kernel density function, PKDF, applied in the analysis[10]. This kernel density model is partially inspired by Sergerie et al. (2021) regarding the way to apply a kernel density model to a granular population of businesses in Canada, however it takes the fundamental structure from statisticians Emanuel Parzen (1962) and Murray Rosenblatt (1956) who independently created the kernel density theoretical form. The notation used in this paper is our very own. This paper is about an applied contribution, with the generation of the cluster heat maps. This paper has no theoretical contribution, outside of the kernel bandwidth design (presented above). This paper uses several existing results from the literature for the explanation of the theoretical kernel density mathematical terms and concepts, and for the explanation of the random processes applied before the kernel density model (explained below), for the perspective of a spatial population of DB polygons and job locations. The remaining terms to be explained below are; i , $n()$, $JW()$, $JL()$, J , $\Theta()$, and $r()$.

$n\left(\Phi, \Phi c, \psi\right)$ represents the finite non-negative number of job locations available within the circular and symmetric surrounding $\pi\psi^2$ (kernel area) of the grid output cell of interest Φ and included in the calculation of the cell total density. Therefore, $n()$ is dependent on the grid output cell id being assessed Φ , the cell geometric centroid location Φc , and the length of the kernel bandwidth ψ . Consequently, the number of DB it covers will vary depending on where Φ is located, and it can cover more than one DB or less than one DB. i is the job index for all unique jobs included in $\pi\psi^2$ around the grid output cell of interest Φ and can only range from 1 to $n()$.

The next few pages document the random processes used in our methodology. This documentation is important to understand the structure of such processes. $JL(i) \sim U(DBP)$ is the job geographic location i for the purpose

10. A normal distribution for the kernel density approach could be explored in future research. However, for this project, we keep the original setting of Sergerie et al. (2021), as it already properly defines each job location contribution to the final density. Furthermore, as it will be described later in this paper, the polynomial form of the kernel density function behaves similar to a Taylor Series Approximation for the normal distribution and replicate quite well the shape of a normal distribution for the domain of interest.

The next few pages document the random processes used in our methodology. This documentation is important to understand the structure of such processes. $\mathrm{JL}(\mathrm{i}) \sim \mathrm{U}(\mathrm{DBP})$ is the job geographic location i for the purpose of the kernel density estimation, it is uniformly and randomly distributed $(\mathrm{U}())$ on its respective DB polygon superficies DBP, and it is not necessarily located at the location of the original establishment the job i belong to. Also, job location i is not allowed to be randomly distributed outside of the boundaries of its own DB, not even within the boundaries of its own DA. Furthermore, all job locations $\mathrm{i's}$, included in $\mathrm{i}=1$ to $\mathrm{i}=\mathrm{n}$ and belonging to same DB, are i.i.d., that is, independent and identically distributed from the same spatial distribution. We call this random process RP.

$\mathrm{JL}()$ is i.i.d. in the sense that a unique job location can be spatially and randomly allocated independently of the spatial random location of another unique job. However, like it will be explained later in this paper, the random process must also be seen as the number of unique jobs randomly allocated on a specific spatial spot of the DB. In this perspective, the larger the number of unique jobs randomly allocated on a specific spatial spot, the less likely another spatial spot will get many jobs, because the number of unique jobs in a DB is finite and not infinite. For this reason, the key insight is to visualize the phenomena as data points (a large set of unique jobs) randomly allocated to the multi-categories (several spatial spots on a DB) of a multinomial random variable (a DB).

Some perspectives about the random process RP are worth mentioning. Even though the data allocation is intended to be uniform, the random process cannot always generate a uniform coverage for a DB, and some neighborhood of the DB could be more covered than others. More specifically, the most likely event of our process RP is a spatially uniform allocation of data points within the DB, where each spatial spot gets the same number of jobs. This is our expected event $\mathrm{E}(\mathrm{RP})$. On the other hand, the most unlikely, but still possible event, is when all distinct data points overlap at the same unique spot in the DB. This is our rare or tail event $\mathrm{T}(\mathrm{RP})$.

Example for a small DB using a uniform binomial random process

To better understand the random process RP of our methodology, an analogy can be provided with a simple uniform binomial random process (BRP) with b unique very small DBs of two spatial spots and v unique jobs per DB, where v is a large and even number, and b is a large number. To build the intuition in a simple manner, we analyse mainly the expected and rarest event of the distribution, and some other events surrounding the expected and rarest event, and not any other details of the distribution. Analogically, an unbiased coin proposes a uniform spatial process made of two spots since both sides of the coin are of equal probability and there is no priority for any of the two sides. However, there is no guarantee that the allocation of head and tail will be even all the time. The most likely event is a perfect mixed bag of heads and tails. That is, the expected event $\mathrm{E}(\mathrm{BRP})$ is a situation where each side of the coin gets an equal allocation of $\mathrm{v}/2$ unique jobs. This situation will happen to a very large proportion of the b DBs. On the other hand, the most unlikely event, but still possible event, is to get all v jobs on one side of the coin only (either head or tail). That is, all v unique jobs get randomly allocated on only one of the two available spatial spots. This situation is our $\mathrm{T}(\mathrm{BRP})$ event, and it will happen to a negligible proportion of the b DBs. Intuitively, all possible combinations of v jobs allocated on the spatial spot of head and tail gets the same probability of realization, but since the event of v head or v tail is one possible combination, respectively, and the event of a perfect mixed bag of head and tail is about a very large number of distinct combinations, then the perfect mixed bag event is more likely to become the expected event.

Formally, we have the following structure for the number of possibilities,

$$\sum_{\mathrm{j}=0}^{\mathrm{v}} \mathrm{C}(\mathrm{v},\mathrm{j}) = \sum_{\mathrm{j}=0}^{\mathrm{v}} \frac{\mathrm{v}!}{(\mathrm{j})!(\mathrm{v}-\mathrm{j})!} > \mathrm{Ec}(\mathrm{BRP}) = \mathrm{C}\left(\mathrm{v},\frac{\mathrm{v}}{2}\right) = \frac{\mathrm{v}!}{\left(\frac{\mathrm{v}}{2}\right)!\left(\mathrm{v}-\frac{\mathrm{v}}{2}\right)!} > \mathrm{Tc}(\mathrm{BRP}) = \mathrm{C}(\mathrm{v},0) = \frac{\mathrm{v}!}{(0)!(\mathrm{v}-0)!} = \mathrm{C}(\mathrm{v},\mathrm{v}) = \frac{\mathrm{v}!}{(\mathrm{v})!(\mathrm{v}-\mathrm{v})!} = 1$$

where $C(v,j) = \binom{v}{j}$, is the number of ways of picking j item among v item without order[11], the right inequality ($>$) is for the rarest events, the middle inequality ($>$) is for the expected event, the left inequality ($>$) is for the total number of possible combinations, and the j index of value 0 and value v is for v tail and v head, respectively.

As a supplementary analysis, let's note that $C\left(v,\frac{v}{2}-1\right) = C\left(v,\frac{v}{2}+1\right)$ represents the number of combinations for events that are one step away from a perfectly balanced spatial allocation, respectively, and are equally less likely to happen than the expected event related to number of combination $C\left(v,\frac{v}{2}\right)$ but still very likely to happen in probability and quite close to the expected event probability. The same way around, $C(v,0+1=1) = C(v,v-1)$ can be seen as the number of combinations for events that are one step forward a more uniform spatial allocation, respectively, and still very unlikely to happen in probability, but quite more likely to happen than the $T(BRP)$ event (Grimaldi, 2003).

We should also visualize the analysis in terms of the summation of equally weighted random variables (non-weighted average). That is, if we label each of the two spatial spots of the DB with numerical values +1 and -1, then for the expected event $E(BRP)$, the number of +1 is equal to the number of -1 and the average is,

$$\left((+1)*(v/2) + (-1)*(v/2)\right)/v = 0$$

If we move one step away from a perfectly balanced spatial allocation, we have an average of,

$$\left((+1)*((v/2) - 1) + (-1)*((v/2) + 1)\right) / v \text{ and } \left((+1)*((v/2) + 1) + (-1)*((v/2) - 1)\right) /v$$

which are different than an average of 0, respectively, but still close to 0, respectively.

For $T(BRP)$, we get either the numerical value -1, v time in a row or the numerical value +1, v time in a row, and the average is,

$$-1 = (-1*v) / v \text{ and } +1 = (+1*v) / v$$

respectively. If we move one step forward a more uniform spatial allocation, we have an average of,

$$\left((-1)*(v-1) + (+1)*(1)\right) / v \text{ and } \left((+1)*(v-1) + (-1)*(1)\right) / v$$

which are different than an average of -1 and +1, respectively, but still close to -1 and +1, respectively.

We now understand that the distribution of DBs is not only about a perfectly uniform spatial allocation for each DB of the CMA and a symmetric and smooth degradation of uniformity is present on both side (left and right) of the average event, if the number of jobs in each DB is large and if the number of DBs in the CMA is large. Repeating the exercise for the whole support of the distribution (and not just the average and rarest events) will generate the shape of an approximated bell curve distribution. It is now time to bridge the existing univariate central limit

11. This paper makes use of the term combinatorial combination and not combinatorial permutation. The former doesn't use the notion of order, but the latter does. This is because we focus on the final spatial allocation of jobs and not the sequential details of the way the random allocation was realized. Also, this analysis focuses on the number of occurrences (or combinations) related to an event of interest (most common and rarest event) and not the probability of each occurrence. That is, for the case of our application, the probability of each occurrence is the same because the random processes are uniform. Therefore, the notion of probability of each occurrence is not relevant for our comparative analysis and the ratio of sum of number of occurrences related to a specific event of interest to the total number of occurrences is enough to get an exact probability perspective (see annex 1 for mathematical proof). Formally, the probability of one occurrence of the rarest event and one occurrence of the average event is the same due to uniformity. That is, $\left(\frac{1}{2}\right)^{v} = \left(\frac{1}{2}\right)^{\frac{v}{2}} * \left(\frac{1}{2}\right)^{v-\frac{v}{2}}$.

theorem for the binomial distribution and its finite variance to our explanation[12]. That is, Asymptotically, if $v \to \infty$, then BRP converge in distribution to the univariate normal distribution $(BRP \to N())$ [13]. In a finite context, if v is large enough, then we can approximate BRP[14] such that[15],

$$BRP \approx N\left(v * 0.5, v * 0.5 * 0.5\right)$$

where the term 0.5 represents the equal chance of allocating a job on one of the two spatial spots of the DB, the $v * 0.5$ term represents the expected number of jobs to be allocated on one of the two spots, which is half of the total number of unique jobs available from all establishments of the DB, and where $v * 0.5 * 0.5$ represents the variance of number of jobs allocated on one of the two spots of the DB (Severini, 2005). Last but not least, in a finite context, we can expect to observe a realised approximated normal distribution over the set of DBs of a large CMA if b is large enough in the CMA and v is large enough in each DB of the CMA, or in a theoretical context, a normal distribution if $b \to \infty$ in the CMA and $v \to \infty$ in each DB. The paragraph closes our binomial random process example.

Uniform multinomial random processes for large DBs

The mathematical calculations for our main random process RP are a similar idea, however, uses a uniform multinomial process (with finite variance) instead of a uniform binomial process (with finite variance) because the number of spatial spots per DB is finite[16] and large, and not only equal to 2. Furthermore, from the existing Multivariate Central Limit Theorem (MCLT) for the multinomial process, if $v \to \infty$ then the convergence in distribution of the uniform multinomial process is a multivariate normal distribution instead of a univariate normal distribution (Severini, 2005) ($RP \to N()$). In a finite context, if v is very large, then the process RP can be approximated such as,

$$RP \approx N\left(vp^T, vM\right)$$

where v is our usual large number of job notation, and p is an S-dimension[17] vector made of S probabilities summing to 1 for the large number of spots S of a typical DB of a CMA. For convenience matter, and without losing too much generality, v is also divisible by S to allow an exact uniform spatial allocation[18]. In the special case of our applications[19],

$$p = \langle \frac{1}{S}, \frac{1}{S}, \ldots, \frac{1}{S}, \frac{1}{S} \rangle$$

and has an equal probability for each of the S element of the vector because the process RP is a spatially uniform multinomial random process and doesn't prioritize any spatial spot among the full set of unique spots of the DB in the CMA. Consequently, vp is the vector of S entries representing the expected number of jobs allocated to each of the S spatial spot of the DB, which is $v * \frac{1}{S}$ for each spot of the DB, because $\left(v * \frac{1}{S}\right) * S = v$. In other words,

$$vp = \langle \frac{v}{S}, \frac{v}{S}, \ldots, \frac{v}{S}, \frac{v}{S} \rangle \text{, and}$$

12. The way we bridge the central limit theorem to the explanation is similar to Rao, 1981 (page 3, section 2.2 asymptotic distribution, paragraph 1)
13. All asymptotic normal distributions of this paper consider a fixed, and properly scaled and centered normal distributions. For simplification matter, the mean and covariance are not displayed in the expressions to not focus the analysis on a mean vector exclusively populated with zeros. Also for simplification matter, random processes for finite and asymptotic context preserve the same notation even though they are not scaled and centered the same way.
14. This paper make us of the notation RP as a random process because, technically, the random allocation of jobs is sequential and a time dimension is present. However, for simplification matter, the analysis focuses solely on the final allocation without considering the sequence.
15. All normal distribution approximation in this paper is presented as un-scaled and un-centered.
16. We assume a finite number of spatial spot available in any DB polygon since the number of pixels is finite and not infinite.
17. All vector dimensions of this paper are initially defined using the 1-row notation (or horizontal vector) for convenience matter in the text. Consequently, the use of the vector transposition refers to a 1-column vector (or vertical vector).
18. v is divisible by finite quantity S for the purpose of the analysis and simplification matter; however, we avoid this simplification for the purpose of the convergence $(v \to \infty)$.
19. It is worth noting that if the randomisation was not exactly uniform and neighborhoods of establishment building's centroids were relatively more prioritized, then vector p would level up terms related to these spatial spots withing these neighborhoods $(> \frac{1}{S})$ and level down all other terms off neighborhoods $(< \frac{1}{S})$, such that the sum of terms is kept to 1.

vM is the variance-covariance matrix of dimension $S \text{ x } S$ of the multivariate normal distribution of our converging uniform multinomial random process. It is expressed in a finite context as,

$$vM = \begin{bmatrix} v*\frac{1}{S}*\frac{S-1}{S} & (-1)*v*\frac{1}{S}*\frac{1}{S} & \cdots & (-1)*v*\frac{1}{S}*\frac{1}{S} \\ (-1)*v*\frac{1}{S}*\frac{1}{S} & v*\frac{1}{S}*\frac{S-1}{S} & \cdots & (-1)*v*\frac{1}{S}*\frac{1}{S} \\ \vdots & \vdots & & \vdots \\ (-1)*v*\frac{1}{S}*\frac{1}{S} & (-1)*v*\frac{1}{S}*\frac{1}{S} & \cdots & v*\frac{1}{S}*\frac{S-1}{S} \end{bmatrix}$$

$$= \begin{bmatrix} \frac{v(S-1)}{S^2} & \frac{-v}{S^2} & \cdots & \frac{-v}{S^2} \\ \frac{-v}{S^2} & \frac{v(S-1)}{S^2} & \cdots & \frac{-v}{S^2} \\ \vdots & \vdots & & \vdots \\ \frac{-v}{S^2} & \frac{-v}{S^2} & \cdots & \frac{v(S-1)}{S^2} \end{bmatrix}$$

and it include only two unique components among the $S \text{ x } S$ non-zero components due to the simplification related to the uniformity. That is, the component on the diagonal and the one off the diagonal. The matrix notation of vM follows the notation of Ericson, 1969 (Appendix, equation A2) for the proper way of notating a matrix with equal terms on the diagonal, equal terms off diagonal and when the number of rows and columns is even and large. $M = P - pp^T$, where $P = I\,p^T$, and where I is the identity matrix of dimension $S \text{ x } S$. That is, P is a diagonal matrix whose diagonal elements are the items of vector p.

$$P = I\,p^T = \begin{bmatrix} 1/S & 0 & \cdots & 0 \\ 0 & 1/S & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 1/S \end{bmatrix}$$

Also,

$$pp^T = \frac{1}{S*S} + \frac{1}{S*S} + \ldots + \frac{1}{S*S} + \frac{1}{S*S} = \frac{S}{S*S} = \frac{1}{S}$$

The variance-covariance matrix vM informs that the variance of job allocation to any spatial spot is $v*\frac{1}{S}*\frac{S-1}{S}$ and the covariance between any two distinct spatial spots of the DB is $(-1)*v*\frac{1}{S}*\frac{1}{S}$. The minus term of the covariance is due to the less likely chance of getting more job counts on one of the spatial spots as the job counts of the other spatial spot increases (Aitkin, 2022).

The multivariate normal distribution of our converging spatially uniform multinomial random process can be expressed in a finite context in the following manner,

$$y' = SVN(x';vp,vM) = (2\pi)^{-\frac{S}{2}} * \det(vM)^{-\frac{1}{2}} * \exp\left(-\frac{1}{2}\,(x'-vp)\,(vM)^{-1}(x'-vp)^T\right) \text{and } x' \approx N(vp^T, vM)$$

All terms of the S-variate[20] normal distribution $SVN(x';vp,vM)$ are already defined above, with the exception of x'[21], which is a row vector of S-dimension, and made of S entries for the respective count of job allocation for the S unique spatial spots of our DB. Even though multivariate and made of vectors and matrices, $SVN(x';vp,vM)$ outputs a scalar value y' and takes the shape of a univariate normal distribution (Aitkin, 2022). $SVN(x';vp,vM)$ reaches its maximum density[22] when $x' = vp$, which is the expected job allocation vector presented above. That is, when $x' = \left(\frac{v}{S},\frac{v}{S},\ldots,\frac{v}{S},\frac{v}{S}\right)$. This statement was pointed out by Thompson and al, 2022 (page 78 (8) equation 2.9) for the general case of a multivariate normal distribution with log transformation. Also, $SVN(x';vp,vM)$ reaches its minimum density for $x' = (v,0,\ldots,0,0)$, which is a situation when the total of unique jobs of the establishments of the DB are allocated exclusively to the first spatial spot only. The same minimum density will be reached for any other single spatial spot of the DB. Finally, $\det(vM)$ or $|vM|$ represent the determinant of the vM variance-covariance matrix. $|vM|$ is a non-zero scalar value, because vM is non-singular and the inverse matrix of vM, $(vM)^{-1}$, does exist. This paragraph completes our explanation about the stability of the random processes involved in our methodology. In other words, we documented the idea that our random processes relate to the normal distribution and the use of random variables converges toward a well centered result and are not a synonym for inaccuracy.

It is worth noting that, since the process RP for job allocation is a random process, then, consequently, any other metric generated on top of this pre-processing is also a random variable. This basically includes all steps of the KDE program. The quantification of uncertainty due to the random pre-processes of our KDE program is not the focus of this paper, but could be the topic of future research. This topic is relevant because the local neighborhood spatial accuracy matters and could vary depending solely on some random realization of rare events or other less uniform events. However, if large number of jobs are available then the magnitude of the uncertainty is expected to be reasonable. Furthermore, as documented above theoretically, if the random processes relate to the normal distribution (that is, Student's t-distribution with degree of freedom $\to \infty$), then most of the random events will be quite centered toward the perfectly balanced expected event and the two tails[23] of the distribution will be thin in density for the rare, and extremely unbalanced events, which is the opposite of what a Cauchy distribution would be (that is, Student's t-distribution with 1 degree of freedom) (Fisher, 1925) and (Hurst, 2010).

Polynomial Kernel Density Function

We now proceed with the remaining parameters of the methodology. $JW(i)$ is the job weight variable for job i, where i is positive and equal or below $n()$, and JW is equal to unity, that is, a value of 1 for all unique jobs. The alternative option would have been the extremum scenario. That is, to spatially allocate a number of job points equal to the number of unique establishments within the DB and select a weight equal to the number of unique jobs within the respective establishment for each job point. This scenario makes better use of the $JW(i)$ weight variable because it is not equal to unity. However, this scenario is not in favor of pre-processing and cleaning the data for the purpose of the KDE, as it makes the spatial discrete distribution of jobs sparser and harder to smooth. The $JW(i) = 1$ setting is preferred because it is about a very large number of equal weighted entities spanning across space and pre-smoothing the original spatial distribution of jobs before applying the KDE.

20. All vectors and matrices dimensions are technically based on $S-1$, and not S. We keep S for simplification matter in the notation. By definition, the last category of a multinomial random variable is a redundant information because the sum of counts per categories is a fixed total constraint, and the count of jobs allocated to the last category is automatically deducted by the sum of all other categories. Consequently, the variance-covariance matrix is rank deficient and not full rank. To avoid singularity, the impossibility of matrix inversion, and improper matrix determinant, the dimension of the multivariate normal distribution needs to collapse within a sub-space of S, which is simply, $S-1$.
21. This paper makes use of variables x, x' and x''. Those are not variation of the same variable. Those are distinct variables used for convenience matter. x is an input ratio in the kernel density function, x is also used as a product symbol for the description of the dimensions of matrices, x' is an input vector for the multivariate normal distribution and x'' is the input variable of a Taylor Series Approximation. y, y' and y'' are related to the same equations, respectively, for the output part.
22. This paper make use of the term density because the distribution is a normal approximation. This paper avoid the term ''mass" for matter of simplification, even though the input vector of the normal is finite and made of the categories of the multinomial.
23. The use of the central limit theorem and normal approximations for multinomials will undeniably be more reliable for statistical inference around the average of the distribution. Nevertheless, we take the time here to mention the characteristics of the tails, as those are convenient compare to the Cauchy distribution.

For a more compact notation of the latest terms introduced in the previous paragraphs, we define the set J of dimension $n \times 2$[24] as $\{(JW(i), JL(i)); i \in (1,2,\ldots, n-1, n)\}$ and we include it as a dependent argument of the $PKDF()$. Furthermore, $\Theta(JL(i), \Phi c, \psi)$ is the Euclidean distance between the grid output cell geometric centroid and the random job location . It is dependent on $JL(i)$, Φc, and the bandwidth ψ. The ratio $1/\psi^2$ is a simple normalizer term outside of the summation and constant for all grid output cells within a CMA and that vary only across CMAs. Finally, if we substitute the two squared exponents with a repetition of their terms, the ratio,

$$r(\Theta, \psi, i) = \left(1 - \left(\frac{\Theta(JL(i), \Phi c, \psi())}{\psi()}\right) * \left(\frac{\Theta(JL(i), \Phi c, \psi())}{\psi()}\right)\right) * \left(1 - \left(\frac{\Theta(JL(i), \Phi c, \psi())}{\psi()}\right) * \left(\frac{\Theta(JL(i), \Phi c, \psi())}{\psi()}\right)\right)$$

is the key term of the kernel density function $PKDF()$. The $r()$ ratio properly allocate a job location i closer to the grid output cell's geometric centroid with a larger individual contribution to the total density of the cell of reference. For its part, a job location i farther from the cell geometric centroid gets a smaller individual contribution. Formally, the polynomial kernel density function generating the total density value y and can be noted as,

$$y = PKDF\left(X, \Phi, \Phi c, \psi(), J(JW(), JL())\right) = \frac{1}{\psi()^2} * \sum_{i=1}^{n(\Phi, \Phi c, \psi)} \left(\frac{3}{\pi} * JW(i) * \left(1 - \left(\frac{\Theta(JL(i), \Phi c, \psi())}{\psi()}\right)^2\right)^2\right)$$

$PKDF()$ is dependent of X, the full CMA spatial job distribution, and not just the subset of job elements of X indexed in i for a specific grid cell of interest Φ. For this paper, to simplify the methodology, X is acknowledged as a fixed distribution, and not subject to the uncertainty related to a super-population model generating a complete random realization of a CMA population, nor to a random or non-random sample aiming at representing the CMA population.

For simplicity matter, we note the ratio $\frac{\Theta()}{\psi()}$, included within the ratio $r()$, as the ratio x. We also note $f(x) = \frac{3}{\pi} * JW(i) * r() = \frac{3}{\pi} r$. A quick function study is essential to understand the shape of $PKDF()$. $f(x)$ can be expressed as $\frac{3}{\pi} - \frac{6}{\pi}x^2 + \frac{3}{\pi}x^4 \approx 0.95 - 1.91x^2 + 0.95x^4$. The first derivative of $f(x)$ is $\frac{df(x)}{dx} = f'(x) = -\frac{12}{\pi}x + \frac{12}{\pi}x^3$ and $f''(x)$ is equal to zero for $x \in \{-1, 0, +1\}$. The second derivative of $f(x)$ is $\frac{d^2 f(x)}{dx^2} = f''(x) = -\frac{12}{\pi} + \frac{36}{\pi}x^2$ and $f''(x)$ equal zero for $x \in \{-\left(\frac{1}{3}\right)^{1/2} = -0.577, +\left(\frac{1}{3}\right)^{1/2} = +0.577\}$. Consequently, $f(x)$ is a decreasing polynomial function for $x \in \{(-\infty, -1) \cup (0, +1)\}$, that is, $df(x)/dx < 0$. $f(x)$ is an increasing polynomial function for $x \in \{(-1, 0) \cup (+1, +\infty)\}$, that is, $df(x)/dx > 0$. $f(x)$ is concave up for $x \in \{(-\infty, -0.577) \cup (+0.577, +\infty)\}$, that is $d^2 f(x)/dx^2 > 0$. $f(x)$ is concave down for $x \in (-0.577, +0.577)$, that is, $d^2 f(x)/dx^2 < 0$. Furthermore, the original domain of $f(x)$ is not limited to a specific subset of the real numbers $\mathbf{R}$. That is, $x \in (-\infty, +\infty)$. The original image of $f(x)$ is the set of the non-

24. This paper make few use of x for the definition of dimensions of items. However, x is mainly used in a mathematical context for variables, vectors and processes.

negative real numbers $\mathbb{R}^+$. That is, $f(x) \in [0,+\infty)$. Finally, even though polynomial, the shape of $f(x)$ is similar to a normal distribution for domain $x \in [-1,+1]$. That is, $f(x)$ is similar to a Taylor Series Approximation (infinite sum of polynomial terms) for a normal distribution in a neighborhood centered at zero. Formally,

$y''(x'') = \frac{1}{\sqrt{2\pi * \sqrt{1}}} e^{\frac{-(x''-0)^2}{2*1}}$, and $x'' \sim N(0,1)$, by definition of a univariate normal distribution of mean 0 and variance 1, and $y''(x'') = \frac{1}{\sqrt{2\pi}} e^{\frac{-x''^2}{2}}$

$= \frac{1}{\sqrt{2\pi}} * \left(\sum_{w=0}^{+\infty} \left((-1)^w * (x'')^{2w} \right) / (2^w * w!) \right)$, by exact equality of experimental Taylor Series result[25], and

$= \frac{1}{\sqrt{2\pi}} * (1 - (\frac{1}{2*1!}) * x''^2 + (\frac{1}{4*2!}) * x''^4 - \ldots)$, by definition of an infinite summation, and

$= \frac{1}{\sqrt{2\pi}} - (\frac{1}{\sqrt{2\pi}} * \frac{1}{2*1!}) * x''^2 + (\frac{1}{\sqrt{2\pi}} * \frac{1}{4*2!}) * x''^4 - \ldots$, by simple distributivity property, and

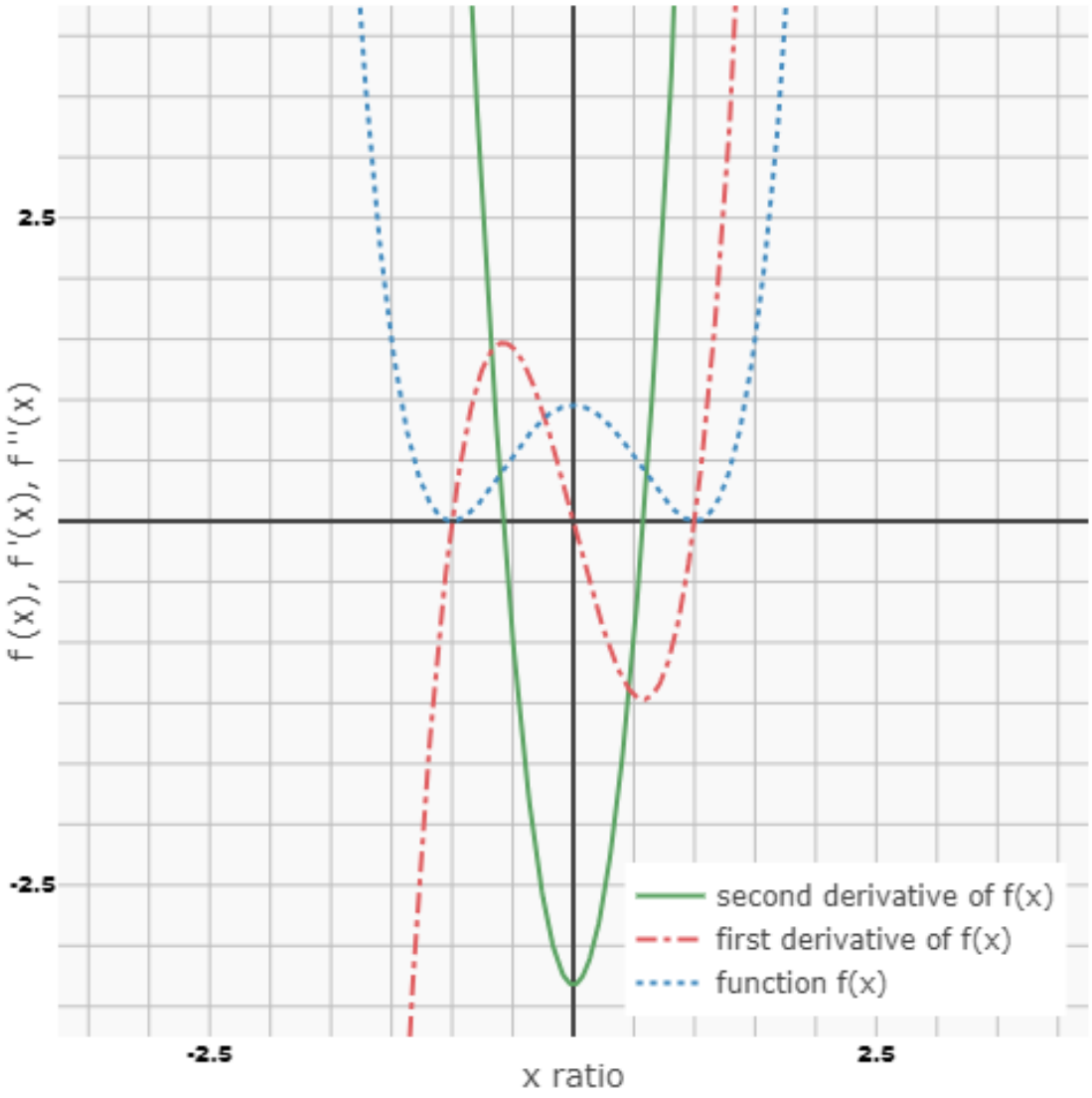
$\approx \frac{1}{\sqrt{2\pi}} - (\frac{1}{\sqrt{2\pi}} * \frac{1}{2*1!}) x''^2 + (\frac{1}{\sqrt{2\pi}} * \frac{1}{4*2!}) x''^4$, after removing an infinity of terms of negligible respective magnitude compared to the first 3 terms, and

$\approx 0.40 - 0.20x''^2 + 0.05x''^4$, after arithmetic simplification and a tough rounding of the π terms

The last approximation above, $0.40 - 0.20x''^2 + 0.05x''^4$, is similar enough[26] to the expression of $f(x) \approx 0.95 - 1.91x^2 + 0.95x^4$ and for 2 reasons: 1) the sign +/- structure is the same, 2) the exponential terms structure is the same. However, the coefficients magnitude structure is not the same. That is, $f(x)$ is quite stable while the other equation is sharply decreasing, which is typical for Taylor Series. Also, the third coefficient of magnitude 0.05 is enough to understand the relative respective negligence of the remaining infinity of removed terms compared to the first 3. The similarities between $y''(x'')$ and $f(x)$ are relevant to understanding the theoretical connection of $f(x)$ with the normal distribution, especially since the normal was an optional parameter in Sergerie et al, 2021 . Figure 3 presents the 3 curves of interest to understand the full picture of $f(x)$, that is, the blue function $f(x)$, its red first derivative function and its green second derivative function. The x-axis is for the ratio x (Bartle and Sherbert, 2011).

25. Taylor Series Approximation for the Normal Distribution (note: this source is experimental but useful to understand our kernel density model)
26. The comparison is made using the scaled expressions, not the un-scaled ones.

Figure 3
Polynomial kernel density function (zoom out for f(x), f'(x) and f''(x))

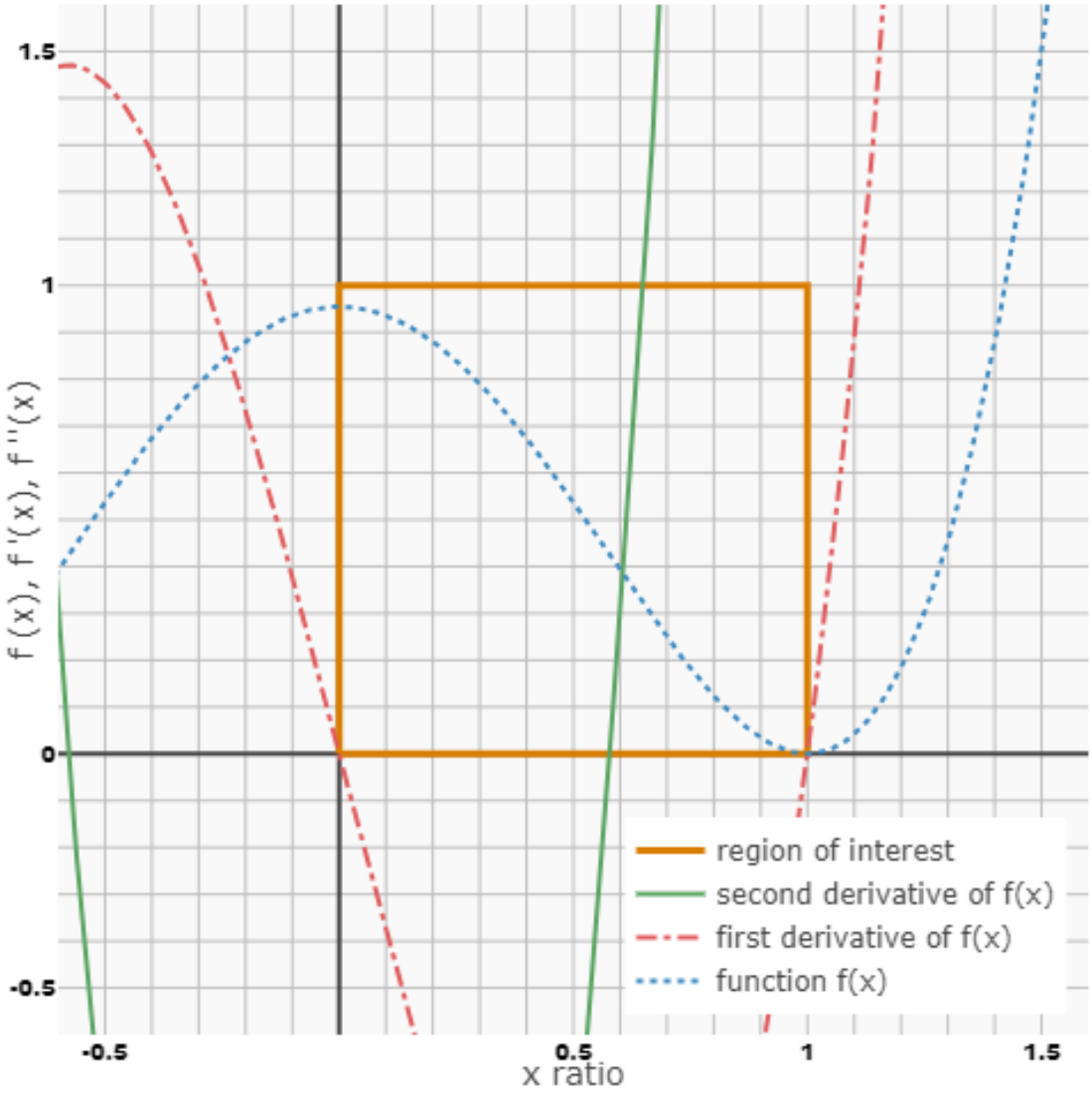


Source: authors' computations from the theoretical kernel density estimation function of Parzen (1962) and Rosenblatt (1956).

Now that the polynomial kernel density function is well described mathematically, we focus on the relevant sub-domain of the function $f(x)$, which is $[0,1]$, and not $(-\infty,+\infty)$. Indeed, the ratio x has no purpose below zero and above 1 in the context of a density estimation exercise, since Θ cannot be below zero (that is, a negative distance from the spatial point of interest is irrelevant) and cannot be beyond ψ (that is, reaching out above the circular area of interest is also irrelevant). Consequently, the relevant image of $f(x)$ is $[0,\frac{3}{\pi}]$, because $f(0)=\frac{3}{\pi}$ and $f(1)=0$ and $f(x)$ is a decreasing function between 0 and 1. More importantly, the distinct contribution of a job location i within the radius of interest ψ is approximately normalized between 0 and 1 (that is, between values 0 and $(3 \ / \ 3.14159) = 0.95)$ and $PKDF()$ is a finite sum of approximately normalized terms, since the number of job locations within the circular and symmetric surrounding $\pi\psi^2$ is finite. Figure 4 is the same as the previous Figure 3. However, it focuses on the relevant domain and image of $f(x)$ and therefore, enable us to better

visualize the value of $f(x)$ when $x = 0$, which is not 1.0, but 0.95. The contribution of job location i decreases as x increases between $[0,+1]$. The decrease is initially less drastic around zero, but eventually becomes sharper and then smoother again just before reaching a ratio x of 1. This structure is due to the second derivative (green curve) reaching a value of zero when x reaches +0.577. A square highlighted yellow bound the region of interest (Bartle and Sherbert, 2011).

Figure 4
Polynomial kernel density function (zoom in at domain [0,1] for f(x), f'(x) and f''(x))



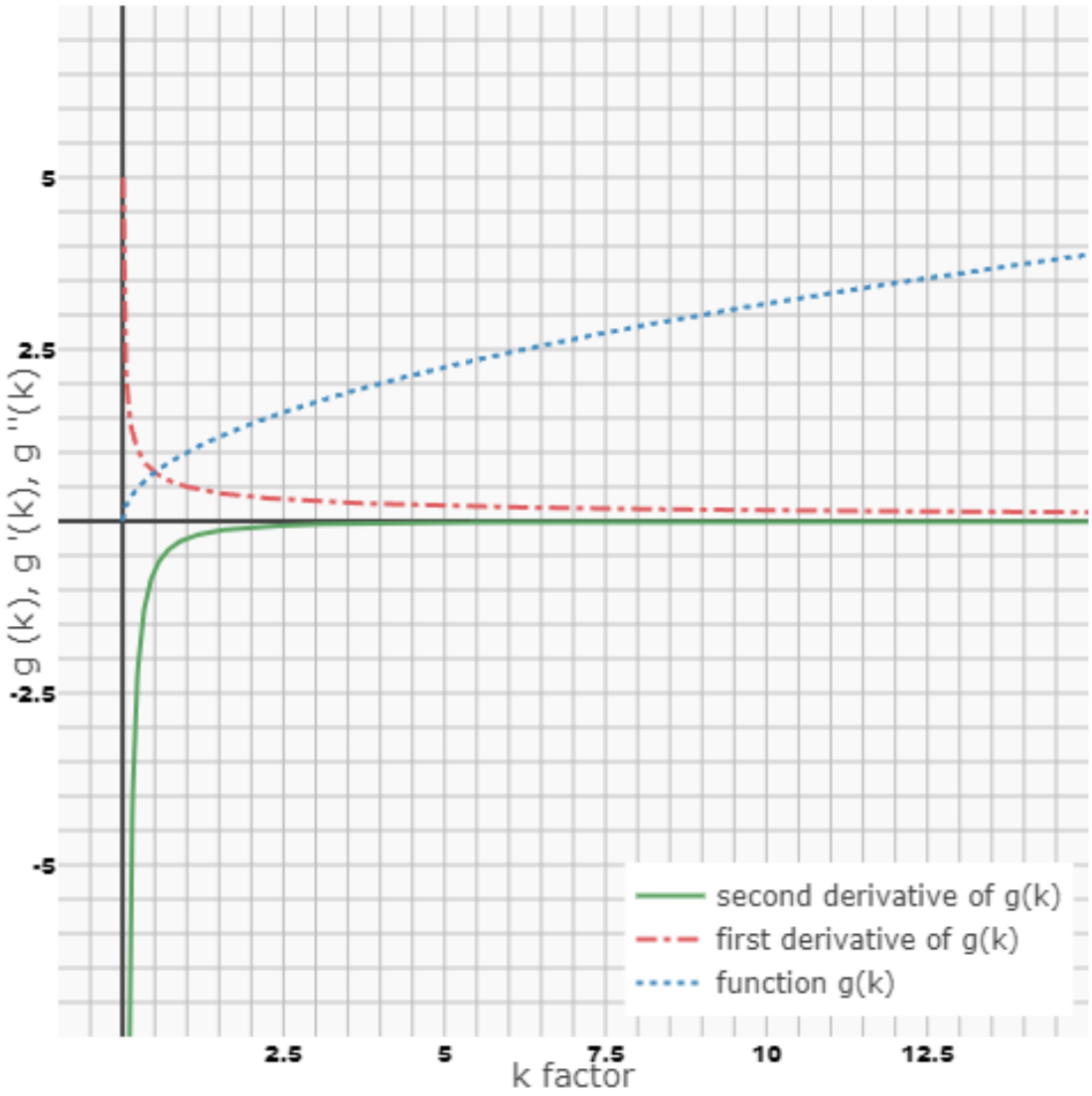
Source: authors' computations from the theoretical kernel density estimation function of Parzen (1962) and Rosenblatt (1956).

Spatial connection between our study and Sergerie et al. (2021)

Before moving to the last part of the methodology where processing steps are done on top of the estimated kernel density distribution for the CMAs and NAICS of interest, we quickly provide an intuitive explanation about the dynamic between spatial density of job locations and spatial distance of job locations. This explanation is essential to properly understand the very nature of our data and its limits. It also provides some connections between our study and Sergerie et al. (2021) in terms of spatial mapping. A paper from the US Census Bureau (Biemer and Stokes, 1984), highlights the following ''Justification: The average distance between randomly distributed points in a plane is increased by $k^{1/2}$ when the density of those points is decreased by a factor of k''. This explanation is intuitively straightforward, however, to better understand their idea, we generate the function $g(k) = k^{1/2}$ and perform a quick study of its first and second derivative properties, domain, and image. In Figure 5, the horizontal axis is the k-axis, and a positive number k represents the decreasing factor in spatial density. The blue curve is the $g(k) = k^{1/2}$ function and it represents the increase in average distance between spatial points. The domain and image of the function $g(k) = k^{1/2}$ is the set of the non-negative real numbers $\mathbb{R}^+$. That is, k and $k^{1/2} \in [0, +\infty)$. The red curve is the first derivative $g'(k) = \frac{d}{dk} k^{1/2}$ and it is equal to $\frac{1}{2} * k^{-\frac{1}{2}}$ and there is no value of k such that it is equal to zero even though it converge to zero, as k tend to infinity $\left(\frac{1}{2} * k^{-\frac{1}{2}} \to 0\right)$. Both its domain and image are $(0, +\infty)$ since no zero can be at the denominator. The green curve is the second derivative $g''(k) = \frac{d^2}{dk^2} k^{1/2}$ and it is equal to $-\frac{1}{4} * k^{-3/2}$ and again, there is no value of k such that it is equal to zero even though it converge to zero, as k tend to infinity $\left(-\frac{1}{4} * k^{-\frac{3}{2}} \to 0\right)$ and, at a faster rate of convergence than the first derivative function due to the larger absolute magnitude of parameters at the denominator (that is $g'(k) = O(k^{-1/2}), g''(k) = O(k^{-3/2})$, and $k^{\frac{1}{2}} < k^{3/2} \Leftrightarrow 1/k^{3/2} < 1/k^{\frac{1}{2}}$). Its domain and image are $(0, +\infty)$ and $(-\infty, 0)$, respectively (Bartle and Sherbert, 2011). Therefore, we explain the justification of (Biemer and Stokes, 1984) in the following way for our own applications about a large population of job locations spatially distributed within a population of DB in a CMA: For a large and fixed number of uniform job locations within the DB of a CMA, as the DB polygon boundaries tend to infinity $(DBP \to \infty)$, and as the spatial density of job's location tend to infinity within the DB $(k \to \infty)$ [27], and as the average spatial distance between job's location tend to infinity $(k^{1/2} \to \infty)$, and as the spatial random uniformity of job locations keep holding within the DB, the variation in average spatial distance between the job's location converge in probability toward 0 between increment. In a finite context, for two distinct large expansions, k and $k + \Delta$, where $\Delta > 0$, and each expansion are implemented independently and from the initial DB polygon boundaries, then $(k + \Delta)^{1/2} - k^{1/2} \approx 0$. The information provided in this paragraph connects the theoretical dots between our study and Sergerie et al. (2021). In other words, the material enables the understanding of how the data of a DB would behave if the surface of randomization expands from a DB to a DA (Dissemination Area) or even an ADA (Aggregate Dissemination Area).

27. Previously in this paper, we assume a finite number of available spatial spots per DB since the number of pixels per DB is finite and not infinite. This was needed to simplify the understanding of a uniform multinomial random process. However, for the theoretical purpose of this explanation above, it is better to assume an infinitesimal perspective for the number of available spatial spots because the jobs should be allowed to spatially move in a continuous and differentiable manner and not a discreet one. That is, the number of spatial spots is infinite and not necessarily countable.

Figure 5
Dynamics between spatial density and spatial distance between points (g(k), g'(k) & g"(k))



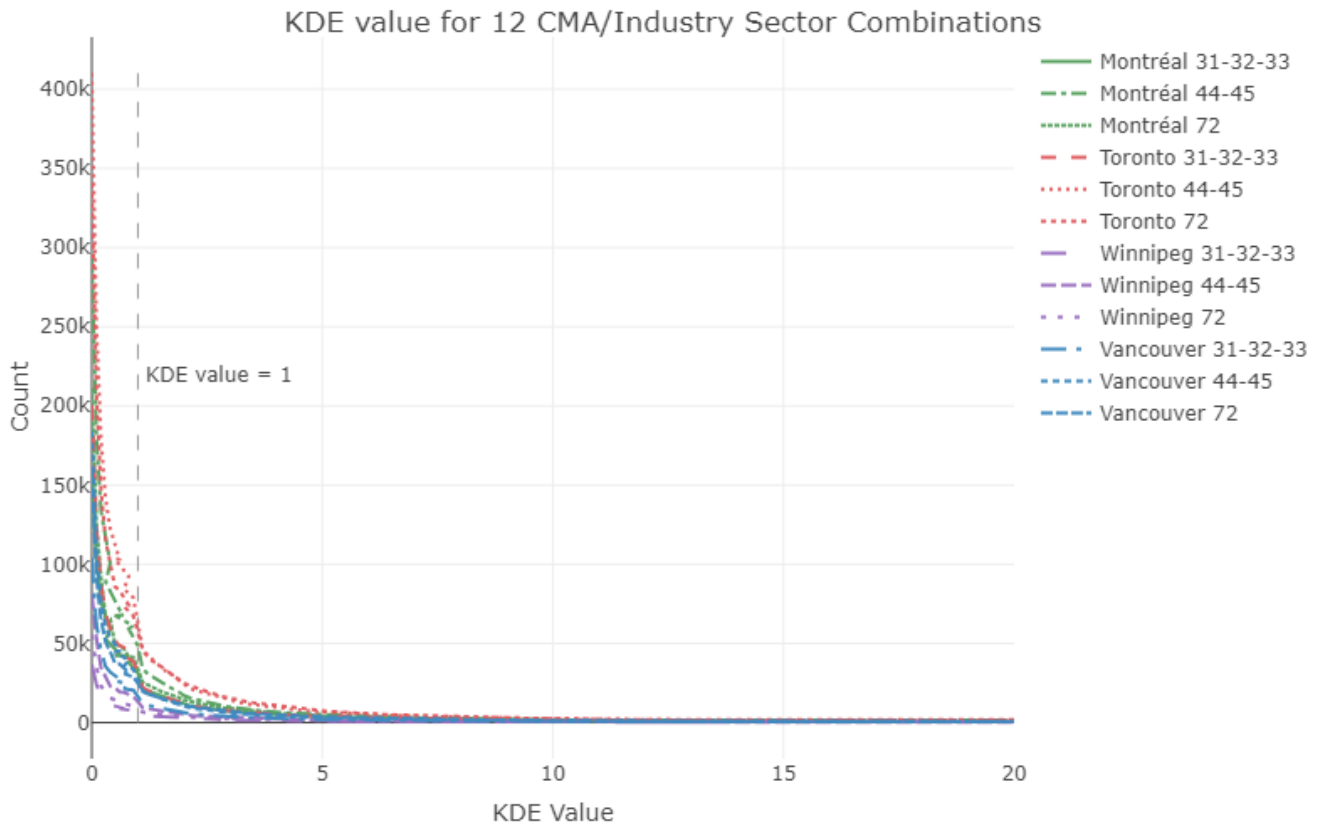
Source: authors' computations based on a textual justification provided in Biemer and Stokes, 1984.

Kernel density thresholds identification

Now, for the remaining part of the main methodology of this paper, using the kernel polynomial density function presented above, a density value for each cell id Φ of the grid is estimated, allowing the generation of an approximately continuous kernel density distribution, as displayed in Figure 6, for three different clusters across four CMAs. In Figure 6, KDE values appear on the horizontal axis and the frequency count in thousands (unweighted number of cells from the full CMA grid) is on the vertical axis. The higher the KDE value, the greater the density of the cell (i.e., the larger the number of job locations available in the cell's neighborhood). As expected, the KDE density distribution is highly skewed. Interestingly enough, the empirical kernel density distribution below, is similar in shape with the theoretical first derivative function of the dynamic between spatial density and spatial distance between points, described in the previous page above (Figure 5). For all CMAs and clusters, the cell counts drop significantly when the KDE exceeds a value of 1 (dashed vertical line on Figure 6). Important drops

are also available for even smaller KDE values, but we ignore those since the KDE values are getting near a density of zero. Given the pattern shown in Figure 6, a threshold of 1 was used as a pre-processing step to filter out the unnecessary segment of the grid cell distribution. However, the analysis was further refined with a second set of thresholds, as described below.

Figure 6
Distributions of KDE values in different CMAs and industry clusters



Source: authors' computations from the BR database.

Using the density distributions displayed in Figure 6, a set of KDE threshold values was tested for each combination of CMA and industry sector by computing the number of DBs, employees, and establishments retained in the clusters at each KDE threshold level. The objective was to iteratively augment the threshold value and remove low-density DBs, and then stop before a significant drop in total employment within the cluster was observed. Ultimately, the KDE threshold values that retained a minimum of about 80% of total CMA employees were applied for most location/industry combinations, which essentially filtered out many DBs populated with small businesses. The KDE threshold values, ranging for the most part between 1 and 3, are shown in Table 2 for each combination of CMA and cluster.

As we can see in Table 2, a large proportion of thresholds preserve the initial value of 1 (Figure 6) and doesn't need an adjustment. This is true especially for the industry clusters defined by Delgado et al. (2014), with one exception. This is an indication that the initial threshold is useful and robust. This result also suggests that the KDE model in this research succeeds to understand the notion of small business neighborhood in Canada, without being ingested with any formal notion of a 3-category classification label system, that is, small, medium, and large business size, beforehand. Indeed, the BR distribution of spatial data point X is continuous and not multinomial. However, 7 out of 24 thresholds were set to a higher KDE cut-off value of 3. Further investigation of these clusters suggests that higher values were more suitable for representing clusters in CMAs with a higher proportion of small

businesses, leading to a more spatially dispersed distribution of employment. In contrast, the threshold value for Hospitality and Tourism in Winnipeg was set at 0.1, indicating a very low concentration of small businesses in this CMA (Table 2). Finally, it is worth noting that the initial setting of thresholds (around value 1) is purely statistics-based and relies only on the shape of the KDE statistical distributions, while the second set of thresholds (Table 2) is driven by economic considerations and the potential for generating meaningful economic indicators from the cluster boundaries. Hence, this composite cut-off setting benefits from a multi-dimensional perspective and is more likely to be stable over time and space. The next three sections finalize the remaining steps of the methodology.

Table 2
KDE thresholds applied to CMA/industry cluster combinations

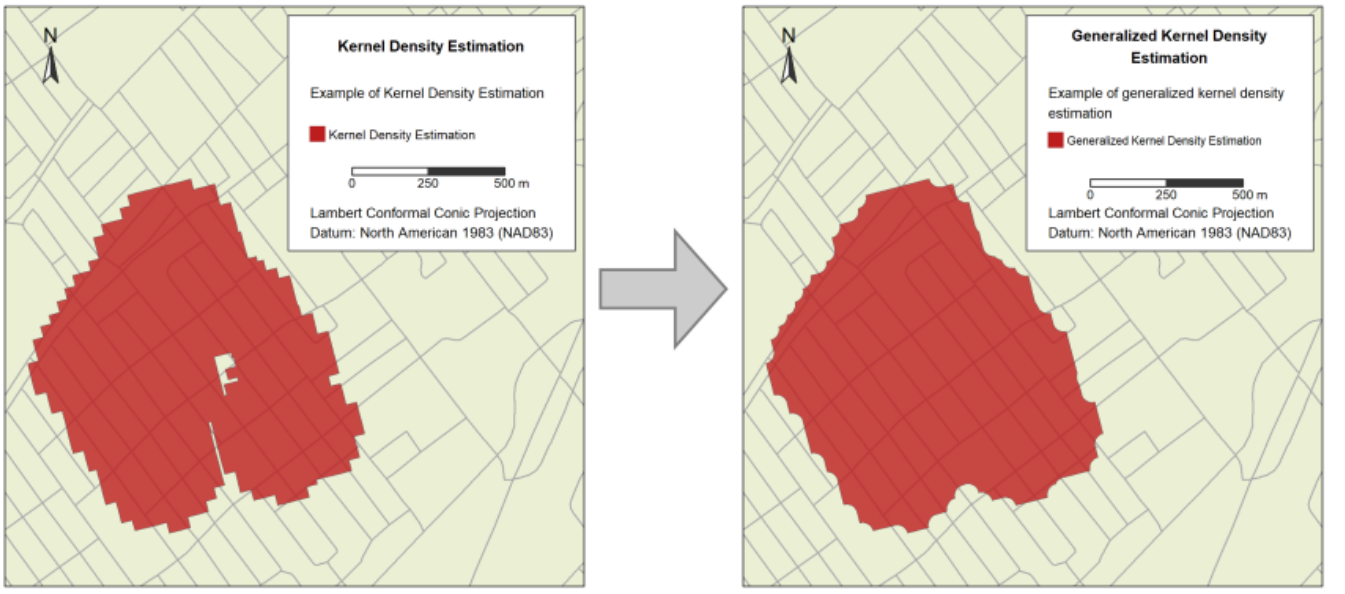
	KDE threshold value			
Industry cluster	**Montreal**	**Toronto**	**Winnipeg**	**Vancouver**
Manufacturing Sector	3	1	1	1
Retail Trade Sector	3	3	1	3
Accommodations and Food Services Sector	3	3	1	3
Distribution and Electronic Commerce (cluster 10)	1	1	1	1
Financial Services (cluster 16)	1	1	1	1
Hospitality and Tourism (cluster 22)	1	1	0.1	1

Source: authors' computations from the BR database.

Generalized KDE

By definition, the KDE results identify areas within a CMA that contain the majority of a specific industry type. Although the filtered results provide an indication of where a specific industry is dominant within a study area, they are not uniform, and small gaps or holes can appear within identified high-density areas. To smooth out these results, a generalization step is applied to the filtered KDE outputs.

Map 1
Changes after generalization step



Sources: Statistics Canada, 2021 Census – Dissemination blocks boundary file, 2021 Census – Population centres boundary file, and authors' computations.

The generalization step begins with the union of all KDE result grid cells into large polygon features. Once the KDE result cells are combined, the polygons are generalized by applying a technique of buffering the union results by 50 meters and then de-buffering by the same amount. This process removes small gaps and holes in the polygons and produces a cleaner output product for conflation. An example of this process is shown in Map 1 above.

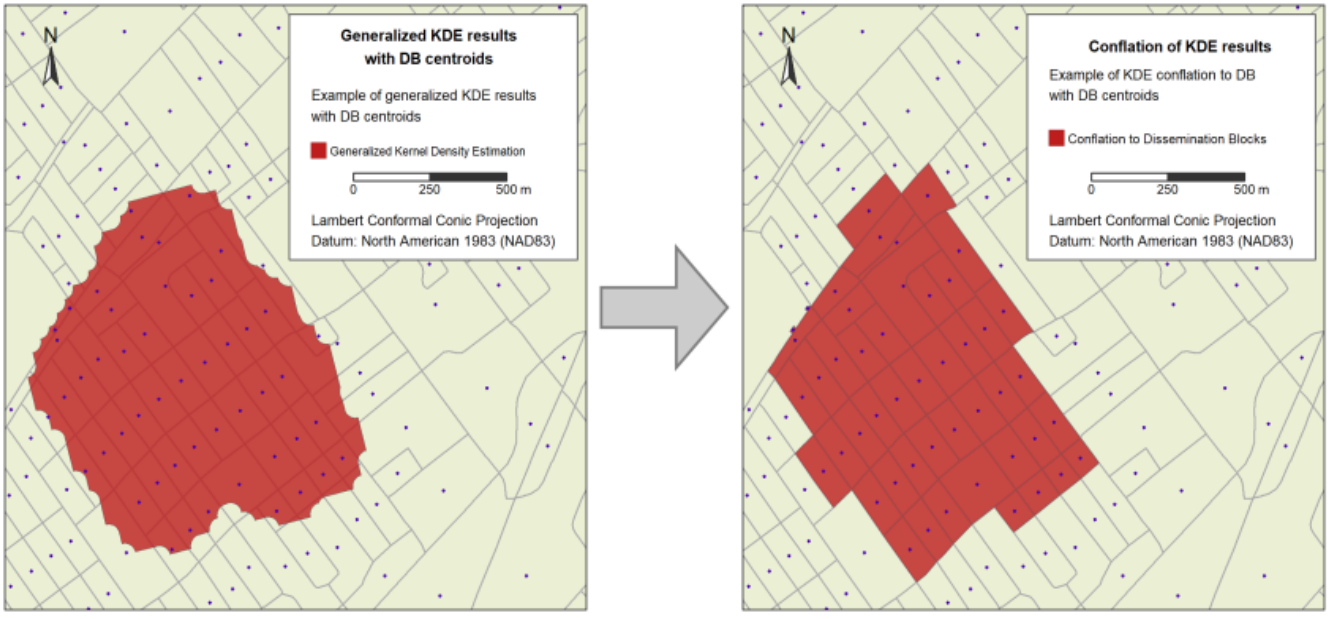
Conflation of KDE results to DBs

The results obtained through KDE and the selection of density thresholds provided an indication of where concentrations of industry clusters exist within each CMA. However, these results were generated at the grid output cell level, and thus lacked association with established boundary files used by Statistics Canada. By re-associating the results with DB boundaries, a more comprehensive analysis of industry clusters can be performed, which is not limited to merely identifying the presence or absence of an industry.

Conflating the KDE results was accomplished by intersecting DB centroids of a CMA with the generalized KDE results. All DBs with centroids that intersected the KDE results were retained to represent industry clusters. In other words, the grid cell overlapping with the DB centroid defines the representative density of the DB. An example of the output from this process (Map 2), illustrates the DB polygons associated with an industry cluster through centroid intersection.

It is worth noting that a weighted DB centroid, based on the BR spatial location of establishment within the DB, could have been used instead of a standard geometric centroid. This is the case, for example, for the Spatial Access Measures Open Database (SAM), engineered at the Data Exploration and Integration Lab (DEIL) of the Center for Special Business Projects (CSBP) of Statistics Canada. However, the methodology of our paper makes use of a uniform randomization process for job location data points within the DB, which, rejects the pertinence of preserving the BR establishment locations within the DB, and validates the use of a simple geometric centroid. As explained earlier in this paper, the expected event $\mathrm{E}(\mathrm{RP})$ of the randomization process RP is a spatially uniform allocation of data points within the DB where each spatial spot gets the same number of data points. Consequently, there is no point in prioritizing a weighted centroid where job locations are likely to be located. Indeed, even though neighborhood accuracy matters, it also must be partially compromised because the objective remains to smooth the discrete spatial distribution of job locations and facilitate the contribution done by the kernel density estimation.

Map 2
Generalization output convert into conflated results through a centroid intersection process



Sources: Statistics Canada, 2021 Census – Dissemination blocks boundary file, 2021 Census – Population centres boundary file, and authors' computations.

Clustering and filtering of conflated results

The results conflated to DB spatial boundaries illustrate which DBs are associated with an industry. However, due to the spatial distribution of some industries, and the irregular shapes of DB polygons, the results of this conflation can generate small groupings of DBs that may be associated with only a few industry establishments. To ensure that the final results focus on the main concentrations of businesses and preserve confidentiality in further analysis, the outputs of the conflation process were grouped into clusters of connected DB polygons, and summary statistics for each polygon were computed.

Clustering of the conflated DB polygons was performed by combining all edge touching DBs. Polygons that only touch at a corner were not considered part of the cluster as shown with the clustering example in Map 3. This rule was implemented to limit the creation of large sprawling industry clusters that have locations scattered across a CMA. With DB clusters identified, counts of employees, establishments, and DB polygons were tabulated for each cluster to assess whether the cluster should be retained.

Map 3
Example of conflated industry cluster results split into clusters of edge-touching DB polygons

Sources: Statistics Canada, 2021 Census – Dissemination blocks boundary file, 2021 Census – Population centres boundary file, and authors' computations.

Using the tabulated cluster results from the previous step, clusters that present a confidentiality risk due to being dominated by a single establishment are filtered out. This was done by removing any clusters that had too few establishments (less than 5) or where too many employees (+80%) by a single business establishment in a cluster. This process highlights the idea that a small cluster isolated from its own CMA largest clusters is not necessarily acknowledged as a confidentiality concern if both confidentiality conditions are met.

Recapitulation of steps from the perspective of a grid cell $\Phi(z)$

Before moving to the next section of this paper, and to stay consistent with our notation, we recapitulate the recent few steps from the perspective of a grid cell of interest $\Phi(z)$. $\Phi(z)$ is initially matched with an estimated density numerical value $u(z)$ included in the non-negative real numbers $\mathbb{R}^+$ (Figure 6). If this value $u(z)$ is below its CMA and industry sector threshold (either a threshold value of 1 or 3 based on table 2), then $\Phi(z)$ is excluded from the rest of the methodology. If $u(z)$ is equal to or above its respective threshold, then $\Phi(z)$ is preserved for the next step. Also, at this point, the $\mathbb{R}^+$ representation of $u(z)$ doesn't matter anymore, and $u(z)$ takes an arbitrary value u equal to the same arbitrary value of all other grid cell $\Phi()$ included in the process of the CMA so far. For the generalization step, if $\Phi(z)$ with value u is included in the process so far, then the generalization also preserves the existence of $\Phi(z)$ and cannot exclude it. That is, the buffering and de-buffering step smooths cluster frontiers

but cannot suppress existing cells of a frontier. However, grid cells that have been excluded so far can now become included as the generalization step is about removing small gaps and holes in the cluster shape. For the conflation to DBs, $\Phi(z)$ is either overlapping with the centroid of its respective DB or not. If not overlapping, then $\Phi(z)$ has no purpose anymore in the process. If overlapping, then the purpose of $\Phi(z)$ is to accept its respective DB into the DB-level representation of the industry cluster heat mapping. Note that we now notate the DB of $\Phi(z)$ as $DB(z)$ because this explanation is from the perspective of $\Phi(z)$ and the DB of $\Phi(z)$ is included in the process so far because of the existence of $\Phi(z)$. At this point, the role of grid cell $\Phi(z)$ is over, and the rest of the decisions are cluster-based, and not cell-based or DB-based. For the final filtering step, the double confidentiality condition will either preserve or suppress the industry cluster where $DB(z)$ and grid cell $\Phi(z)$ are so far included. If the cluster is suppressed, then the contribution of grid cell $\Phi(z)$, no matter how significant in the previous steps, now becomes zero because $DB(z)$ is no longer part of a cluster. If the cluster is preserved, then the grid cell $\Phi(z)$ preserves its contribution because $\Phi(z)$ is the cell whose role was to add $DB(z)$ within a cluster part of the final heat map.

Non-technical summary of the methodology

This methodology section concludes with a non-technical summary of the steps for the generation of the industry clusters. We decompose the approach into 4 main steps: 1) data, 2) kernel, 3) thresholds, and 4) post kernel processing.

Data:

- Extract establishment level data from internal Statistics Canada BR database
- Define location of each unique job part of an establishment (Randomize spatially all job locations of a DB within its own DB if the establishment is located within the DB)
- Define weight of each job (in our case, a weight of 1 for each job)

Kernel:

- Define grid cell shape and dimension for kernel
- Define functional form for kernel
- Define bandwidth length for kernel
- Define geometric centroid of each cell of the grid within CMA
- Compute a unique density value for each cell of the grid within CMA using distance between cell geometric centroid, and job location

Thresholds:

- Calculate cut-offs based on statistical distribution of KDE for each CMA and industry
- Calculate cut-offs based on retention ratio approach for each CMA and industry
- Prioritize cut-offs based on retention ratio approach if different from cut-offs based on statistical distribution of KDE

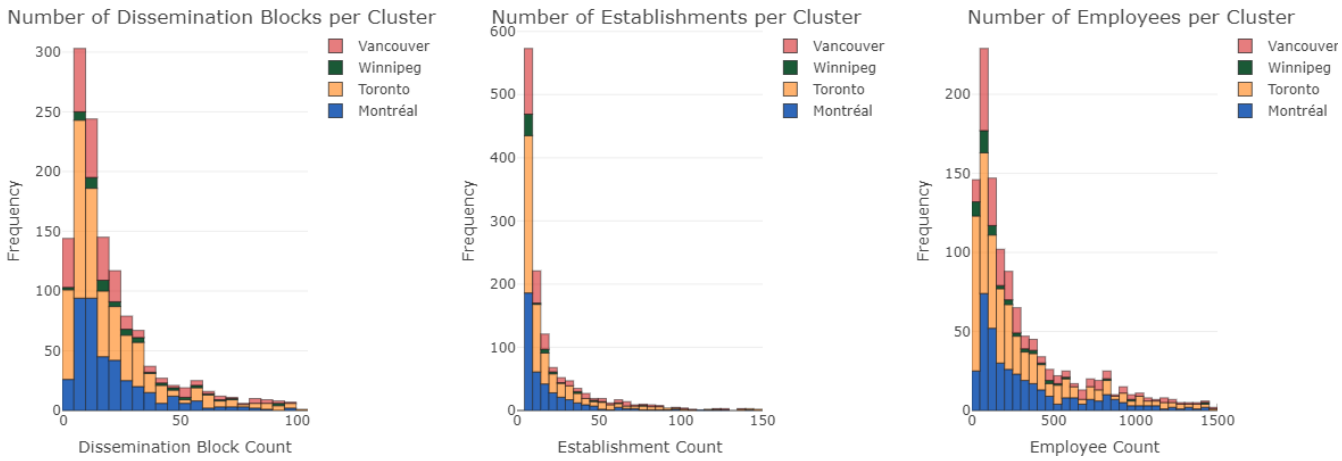
Post kernel processing:

- Generalize KDE by smoothing cluster boundaries and filling cluster inner gaps
- Conflate KDE results to DBs to get a final DB representation of the clusters
- Filter out clusters if not meeting the double condition confidentiality requirement

Results

Figure 7 presents an overview of the descriptive statistics for the cluster results. Number of DBs, establishments and employees per cluster are available for each of the 4 CMA study areas. The x-axis is for the count of DBs, establishments and employees, respectively. The y-axis is the frequency or the count of distinct clusters. Statistical distributions are similar in shape with the highly right-skewed KDE distribution of Figure 6. That is, similar to a power law distribution, which is in line with (Gabaix, 1999) that city size is power law distributed. Furthermore, for all 3 histograms of Figure 7, the frequency is more conservative for Montreal (blue highlighted histograms), which happens to be the CMA with the smallest and lowest range of DB superficies of Figure 1 and one of the lowest KDE distributions of Figure 6 with manufacturing NAICS 31-32-33. Based on these empirical observations, the number of clusters per CMA seems to be correlated with the DB population structure of the CMA. This is an intuitive result. That is, CMAs with a high proportion of large DBs would contribute to generating sparser segments of high-density grid output cells, and consequently produce a larger number of distinct spatial clusters. Alternatively, CMAs largely dominated with small squared DBs would tend to aggregate several clusters into a single large cluster during the generalization, buffering-de-buffering, and conflation process, due to the proximity of DBs to each other. This raises some interesting research questions, and we take the time to describe two here. 1) Does the DB configuration of a CMA explain a significant proportion of the population of clusters? 2) Are there other more important factors affecting the way the clusters and shapes emerge from the spatial data?

Figure 7
Distribution of number of DBs, establishments, and employees for each CMA study area



Source: authors' computations from the BR database.

Key results of the spatial clusters for Montreal, Toronto, Winnipeg, and Vancouver along with six specifications of industry clusters are summarized in Table 3, while the mapping of each cluster is reported in the Appendix. In each map, the clusters for the reference industry sector are highlighted in red.

As expected, spatial distribution of clusters varies substantially among industries, with the Retail Trade and Accommodations and Food Services sectors generally covering the largest shares of DBs in each metropolitan area. This is also reflected in the count of DBs that belong to each cluster (Table 3). For all the CMAs, Retail Trade and Accommodations and Food Services sectors account for the largest number of DBs. Similarly, the DB retention ratio, that is the percentage of DBs that were included in the corresponding cluster in each CMA, ranges from 36% to 55%, and is generally higher (in the vicinity or above 40%) for Retail Trade and Accommodations and Food Services sectors (Table 3).

The clusters capture most of the establishments and, even more significantly, the majority of employment in the corresponding industries (Table 3). As described in the previous sections, this was a key criterion for determining the density value threshold for the inclusion of DBs in the cluster. The establishment retention ratio represents the share of establishments within the industry cluster DBs relative to the total count of establishments in that industry within the CMA; similarly, the employment ratio represents the share of employment generated by businesses within the industry cluster areas relative to the total employment in that industry within the CMA.

Table 3
Results: cluster summary statistics, 2023

	DB	Establishment	DB	Employment	Establishment
CMA Cluster	count		retention ratio		
Montreal					
Manufacturing Sector	3,775	4,136	26.1	89.7	70.9
Retail Trade Sector	9,861	10,431	38.9	92.0	76.7
Accommodations and Food Services Sector	7,888	6,806	36.5	85.4	77.6
Distribution and Electronic Commerce (cluster 10)	5,660	5,406	34.4	94.9	74.6
Financial Services (cluster 16)	2,902	1,512	26.2	88.5	55.4
Hospitality and Tourism (cluster 22)	2,472	860	23.6	60.3	47.1
Toronto					
Manufacturing Sector	5,732	7,594	34.8	94.4	81.6
Retail Trade Sector	9,716	15,075	35.5	97.1	76.0
Accommodations and Food Services Sector	9,946	10,085	38.5	89.2	77.2
Distribution and Electronic Commerce (cluster 10)	7,045	10,291	36.2	97.8	81.9
Financial Services (cluster 16)	5,972	4,856	35.6	95.7	72.0
Hospitality and Tourism (cluster 22)	2,842	1,329	25.3	71.2	50.9
Winnipeg					
Manufacturing Sector	1,054	616	44.2	91.9	78.5
Retail Trade Sector	2,700	2,130	55.3	94.6	87.2
Accommodations and Food Services Sector	2,217	1,313	54.9	91.2	87.7
Distribution and Electronic Commerce (cluster 10)	855	850	33.1	93.2	73.5
Financial Services (cluster 16)	666	349	28.4	82.7	55.2
Hospitality and Tourism (cluster 22)	1,363	337	60.3	81.3	75.2
Vancouver					
Manufacturing Sector	2,432	3,236	35.9	90.8	79.1
Retail Trade Sector	3,981	7,260	36.7	81.3	77.5
Accommodations and Food Services Sector	4,686	5,624	43.7	91.6	83.0
Distribution and Electronic Commerce (cluster 10)	2,662	4,610	32.0	93.7	77.1
Financial Services (cluster 16)	2,166	2,130	32.3	91.2	68.2
Hospitality and Tourism (cluster 22)	1,709	1,041	31.1	79.8	60.1

Source: authors' computations from the BR database.

Except for Hospitality and Tourism (cluster 22), all other industry clusters capture well over 80% of the employment of that industry within their respective industries in the CMAs of reference, with some clusters capturing 95% or more of the employment (Table 3). For instance, the manufacturing sector cluster contains 89.7% of total manufacturing employment in Montreal, 94.4% in Toronto, 91.9% in Winnipeg, and 90.8% in Vancouver.

Similar percentages are computed for counts of businesses that are within the cluster areas. Although these percentages are slightly smaller than those for employment, they remain around 80% for most clusters, except Hospitality and Tourism (cluster 22) and Financial Services (cluster 16). For example, the manufacturing sector clusters contain 70.9% in Montreal, 81.6% in Toronto, 78.5% in Winnipeg, and 79.1% in Vancouver of total manufacturing establishments. This partition of businesses, within the cluster versus those operating outside the cluster, could also be used to monitor trends in the evolution of metropolitan clusters.

It should be noted that, for some clusters, the count of establishments is significantly smaller than the count of DBs comprising the cluster. For instance, the Hospitality and Tourism (cluster 22) cluster in Montreal includes 860 establishments and 2,472 DBs. This result is driven by the buffering and conflation methods used in the analysis and the concentration of businesses in areas with small DBs. The methodological approach developed in this analysis is designed to provide a neighborhood level representation, as opposed to representation of individual DBs. Thus, businesses within DBs that are in proximity to each other, but still separated by other DBs, are clustered together including DBs that are surrounded by such businesses but do not contain any within their boundaries. The opposite situation is also possible. That is, the DB include establishments and job locations but is not part of the final cluster. This can happen, if the grid cell overlapping with our DB centroid is not dense enough and gets filtered out during the 2-stage threshold process explained earlier. Such DB centroid is usually located at the boundaries of a large cluster and get rejected by the process due to the bandwidth aggregating a too large proportion of DB with no establishment location. A quick analogy would be fitting a non-linear regression model over an assembly of data points. The fitted model is sometimes above or below the actual data points, but overall, contributes to provide a good and continuous approximation of the data phenomena, and provides an anonymized representation of the actual spatial distribution of establishments and job locations. The fitted curve enables the analyst to do research about trends and patterns without directly observing the confidential data.

It is worth noting that a higher retention ratio (closer to 100%) for a CMA is not equivalent to a higher level of economic performance of the CMA. Retention ratios are rather closer, in meaning, to the distribution of business sizes, where lower ratios imply larger proportion of small firms within the CMA. The distribution of business sizes is publicly available and not an exclusive information provided by this research but remains a source of validation for the results of this research.

The results obtained with the mapping of business clusters were visually validated by comparing them with land zoning maps of municipalities included in the CMAs of this study. This validation was particularly feasible for manufacturing clusters and industrial zones or industrial parks. For example, for the municipality of Toronto, the industrial zoning of land parcels aligns closely with the location of manufacturing clusters. The use of data from third party sources, such as OSM or Google Maps, was not included in the current analysis; however, it could be considered for future validation efforts.

A few more notes for the understanding of our retention ratio. The retention threshold is based on a simple marginal gain principle. That is, we consider it worthwhile to keep filtering out DBs if the removal is sufficiently larger than the removal of employment. In other words, neatness and visual clarity matters for cluster heat map representation, until it affects economic performance. For cluster 22 Montreal, the retention ratio is 60% of employment. Consequently, choosing 80% of employment was not a motivation because a significant proportion of DB kept dropping at the expense of a small enough proportion of job locations. However, going beyond 60% of employment would not only make the job drop considerably larger than the DB drop, but it would also reach a state where the marginal decrease of jobs is exponential, say, from 60% to 30%, which would be unacceptable. This exponential drop is intuitive and explainable considering the very nature of our statistical BR distributions. As discussed in the previous sections of this report, the BR data are highly skewed with Power Law related shapes. Therefore, the non-flat segment of the distribution includes a chaotic segment, where a very small variation of information leads to a very large drop. Clusters defined over a relatively smaller proportion of establishments are likely to be smaller in cluster size and number of clusters. Investigating cluster 22 for Montreal and Toronto in the annex of this report, the clusters are indeed smaller, more sparse, and less voluminous in number.

Table 4
Percentage of co-location between 2 types of industry cluster for each CMA study area, 2023

	Percent of Co-Location Among Industry Clusters		
	Manufacturing	Retail Trade	Accommodations and Food
CMA Cluster	percent		
Montréal			
Manufacturing	-	37.3	20.5
Retail Trade	30.6	-	43.6
Accommodations and Food Services	24.6	63.6	-
Toronto			
Manufacturing	-	50.1	37.0
Retail Trade	47.5	-	52.7
Accommodations and Food Services	39.9	59.9	-
Winnipeg			
Manufacturing	-	47.5	37.4
Retail Trade	24.6	-	46.0
Accommodations and Food Services	25.4	60.4	-
Vancouver			
Manufacturing	-	39.5	30.8
Retail Trade	46.6	-	61.1
Accommodations and Food Services	30.3	51.0	-

Source: authors' computations from the BR database.

Finally, a simple co-location analysis of business clusters was performed by overlapping boundary files of two clusters and computing the percentage of area of one cluster (row) that is also included in another cluster (column). Table 4 shows the results for an assessment of co-location of Manufacturing, Retail Trade and Accommodations and Food services. A visualization of this co-location between two sets of clusters is also reported in Map 4.

Two key elements emerge from these examples. First, the percentage of co-location varies between approximately 20% and 60% of the cluster area, depending on the specific clusters. However, the strength of co-location between business clusters provides meaningful economic insights. In particular, across all the CMAs analyzed,

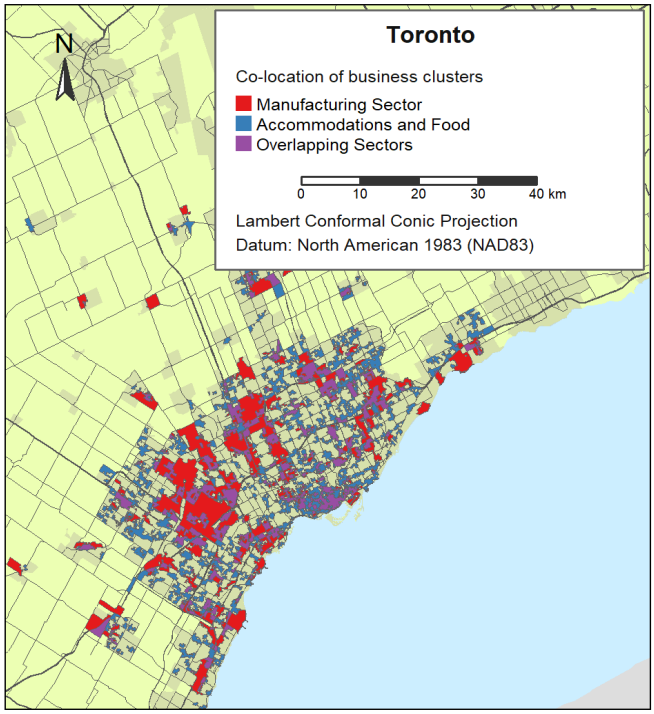
the overlap between retail trade, and accommodations and food services is more pronounced compared to either of these two industries and manufacturing. This observation aligns with the general perception that these types of services tend to co-locate within the same geographic area within a metropolitan area.

Second, the type of co-location provides insights on potential differences that the same industry cluster may present across the metropolitan area. As an example, Map 4 shows that the Accommodations and Food Services sectors are concentrated in neighborhoods that also have a high concentration of retail trade, and separately, a high concentration of manufacturing sector neighbourhood. It is likely that these areas of overlap indicate Accommodations and Food service cater to different clienteles and present different types of sector linkages or dependencies.

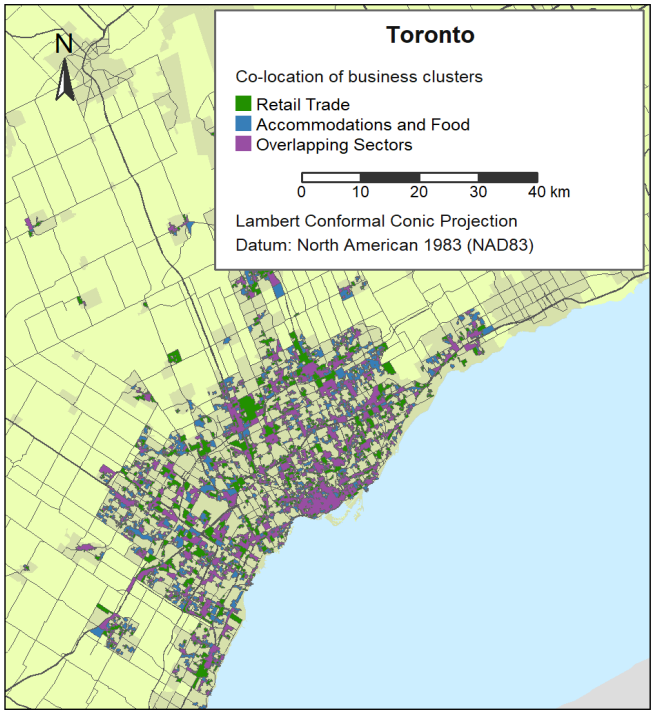
Finally, it should be noted that the concentration of certain types of businesses, like manufacturing, in areas that are outside of what would appear to be their designated municipal zoning areas, such as industrial zoning, may indicate the concentration of specific business functions in commercial areas. For instance, the clustering of manufacturing businesses in downtown Toronto, in areas designated as commercial zoning, suggests that the manufacturing businesses in this area may be related to headquarters or office functions, as opposed to production establishments.

Map 4
An illustrative example of co-location of clusters in Toronto

Panel A: Manufacturing and Accommodations and Food Services Sector



Panel B: Retail Trade and Accommodations and Food Services Sector



Sources: Statistics Canada, 2021 Census – Provinces/territories boundary file, 2021 Census – Population centres boundary file, 2021 Census – Road network file, 2016 Census - Coastal waters boundary file, 2016 Census - Lakes and Rivers boundary file, and authors' computations from the BR database.

Directions for further research, analysis, and applications

The methodology presented in this paper can be further developed and applied to generate insights on business conditions at the local level. The most immediate development involves scaling the work to all metropolitan areas of Canada. Further extension to medium-sized agglomerations (Census Agglomerations) and rural and small-town areas should also be considered.

Increasing geographic coverage can be done in parallel with further refinement of the NAICS groupings that define each cluster. To develop the methodology, the focus of this paper was on simple 2-digit NAICS codes or pre-existing NAICS groupings, with specific reference to the work of Delgado et al. (2014) on U.S. industry clusters. Although there is validity in continuing with this approach, the use of BR microdata provides the opportunity to implement alternative aggregations of NAICS codes and different digit levels. Custom-specific groupings could be considered. For instance, some of the existing literature has focused on artistic and cultural clusters and neighborhood vitality; hence, custom specific definitions of these types of clusters could be considered for implementation.

Further analysis should consider profiling business cluster performance with additional BR variables. In addition to the 6-digit NAICS code, number of employees, and latitude and longitude position for each business establishment, the BR provides several additional fields of interest, which could be integrated into the analysis for profiling and indexing of business performances. In particular, the use of business revenue, total expense, total assets, and date of birth of the business could be explored to generate aggregate spatial indices. Engineering financial performance indicator ratios from the BR could also be possible. There are several of them; 1) profitability ratio ((total revenue minus total expenses) / total assets), 2) liquidity ratio (current total assets / current total debt), 3) tangibility ratio (total fixed assets / total assets), 4) total sales growth rate. It is important to emphasize that BR data comes with technical challenges. The extremely high level of granularity, both in terms of geography and industry group, would limit the type of information that could be extracted. However, the use of indices (e.g., heat maps) or categorical groups (e.g., simple classifications into high, medium, and low values) could mitigate these issues, while providing valuable insights on performance and trends of business clusters at the neighborhood level. This classification would involve clustering highly skewed statistical BR distributions where most of the observations are located within a small range of low values.

On the other hand, the BR remains a very important dataset for Statistics Canada and Canadians. As noted on the Statistics Canada's BR official web page, *''As a statistical register, it provides listings of units and related attributes required for survey sampling frames, data integration, stratification, and business demographic statistics. The BR is a major pillar of the agency's economic statistics programs, including the Census of Agriculture.''*

The BR is maintained on a continuous basis, with inclusion of new businesses and with business employment and financial information updated at regular intervals. New vintages of the BR, with business counts by employment size, are released twice a year; hence, some statistics of cluster maps could be updated at regular intervals.

Finally, the results generated with neighborhood-level clusters could be combined with other data sources, including both Statistics Canada's data holdings and alternative data sources from external providers. As an example, cluster mapping at the neighborhood level could be overlaid with proximity measures to services and amenities and spatial access measures to understand how the presence of amenities or accessibility interacts with the clustering of specific businesses.[28] Similarly, cluster boundaries can be overlaid with mobility flows or commodity flows at a similar geographic scale. This information may provide insights into the level of economic activity within each cluster. Specifically, the combination of cluster boundary files with data on mobility flows is an example of data integration that may yield significant business insights. In this case, the boundary files would be used as geofences to estimate inward-outward mobility, using mobile device data, or mobility through the road network of cluster areas. This information could be used to estimate or monitor the economic activities in the business clusters. Spatial econometric models would be needed to acknowledge the spatial dependencies across these datasets.

As a final note, we observed a similarity of results between Figures 1, 6, and 10, which seems to suggest that the number of clusters per CMA is correlated with the DB population structure of the CMA. Interesting research

28. See: Proximity Measures Database and Spatial Access Measures.

question arises: to what extend can the DB configuration of a CMA explain the way clusters emerge and shape, and what would be the remaining factor explaining the phenomena? Furthermore, the current BR year assessed in this research could be compared with previous BR years available in the Statistics Canada's Database. This analysis would be beyond a simple co-location analysis and exploit the signal located within the space-time BR. Spatial regression and autocorrelation (Moran' I) between several versions of the BR, at neighborhood level, could robustly measure the magnitude of change in the shape of the clusters across years. More specifically, we could identify neighborhoods of a CMA where the current clustered industrial expansion is explained by the spatial historical foundation of other industry clusters. The same for industrial stagnation and contraction at the neighborhood level. If the comparison of several years of the BR is a difficult exercise, then a methodology would need to be formalized following the steps of Statistics Canada methodologies for gross flow statistics and the approaches of sampling across time for the creation of unbiased and consistent statistics.

Also, the current kernel bandwidth acknowledges the DB distribution and the median DB dimension of the CMA but doesn't consider the dispersion of job location. Therefore, a direct improvement would be a bandwidth including the 2 dimensions of information into the calculation. That is, a composite bandwidth weighting 2 components, our approach and the traditional Silverman's approach.

Finally, granular spatial heat mapping engineering, like time series analysis, needs to be validated for the presence of random walks and spurious correlation. For time series, the correlation of 2 series with non-stationary statistical moment, such as the average and variance, can lead to correlations that appear strong but are unlikely to persist after simple transformation like a first time-step lag differentiation. The same applies for spatial data. Two granular spatial heat maps can be non-stationary, get a very strong spatial correlation, or in some cases, a very strong overlap or co-location of clusters, and not preserve the strong correlation after simple spatial transformation, such as spatial lag differentiation. Modern methods to identify robust spatial unit-root and spurious correlation can be found in Muller and Watson, 2023, and Hassan, 2012.

Conclusions

This paper presents an approach to developing business clusters at the neighborhood level using BR establishment-level data for four CMAs in Canada. A new approach to define the kernel bandwidth is detailed since the traditional Silverman's rule bandwidth method fails, in the case of our applications, to directly recognize the configuration of the DB structure within the cities of interest. The generated business clusters fill a data gap on business cluster analysis at a highly granular level of geography, providing a framework that can support decision-making and policy at the local level. It also offers a comparative framework for analyzing trends across metropolitan areas in Canada.

The existing literature suggests that business location choices extend beyond the selection of metropolitan areas; neighborhood characteristics are also relevant determinants of location choices. As a result, business trends can vary greatly across neighborhoods within the same metropolitan areas. In turn, the clustering of businesses in a neighborhood may influence the overall economic prosperity and quality of life in that neighborhood, contributing to either expanding or reducing spatial disparities across the metropolitan area.

The methodology proposed in this analysis draws from existing literature on the identification of central districts. In simple terms, it transforms the discrete and fragmented spatial distribution of job locations (based on the geolocation of establishments) into a relatively smoother or more continuous representation using a fine-level grid to support spatial KDE, and subsequently conflates the results at the DB level. The data for this analysis comes from business records from Statistics Canada's BR. The model introduced a new bandwidth approach to acknowledge more efficiently the configuration of DB within a CMA and be more robust to outliers present in the highly skewed distribution of BR's job count per establishment. The results generated through this process are filtered to remove single DBs or small aggregations of DBs that would not meet basic confidentiality thresholds for business data. The results show that the proposed method is effective in capturing the large majority, if not nearly the totality, of employment in the respective industries, while simultaneously filtering out a large proportion of DB where the density of job locations is relatively weaker and less interesting. Moreover, while the focus of this analysis has been on the delineation of the boundaries (hence, the cluster areas are monochrome in the appendix maps), future analytical applications can highlight different trends with the cluster areas by using, for instance, heat maps.

As the demand for a more granular spatial information on business continues to grow, the use of business clusters at neighborhood level, as spatial framework of reference, can support the work of local economic agents and policy stakeholders. Neighborhood-level analysis can support local business associations interested in understanding and monitoring local businesses within a specific neighborhood. In addition, local cluster boundaries could be integrated with other measures of spatial distribution at the DB-level, such as population density and proximity to amenities, to generate more comprehensive analysis of local conditions and development opportunities.

References

Aitkin, M. (2022). Introduction to Statistical Modelling and Inference. The multinomial distribution. 1st Edition. Chapman and Hall/CRC.

Bekar, C., & Lipsey, R. G. (2001). Clusters and economic policy. *Canadian Journal of Policy Research, 3*(1), 62–70.

Bartle, R. G, & Sherbert, D. R. (2001). Introduction to real analysis. Wiley; 4th edition.

Bathelt, H., & Li, P.-F. (2014). Global cluster networks: Foreign direct investment flows from Canada to China. *Journal of Economic Geography, 14*(1), 45–71. https://doi.org/10.1093/jeg/lbt020

Biemer, Paul. P., & Strokes, S, Lynne. (1984). An improved procedure for the estimating the components of response variance in complex surveys. Bureau of the Census. Statistical Research Division Report Series. SRD Research Report Number: CENSUS/SRD/RR-84/06. https://www.census.gov/content/dam/Census/library/working-papers/1984/adrm/rr84-06.pdf

Campaniaris, C., Hayes, S., Jeffrey, M., & Murray, R. (2011). The applicability of cluster theory to Canada's small and medium-sized apparel companies. *Journal of Fashion Marketing and Management: An International Journal, 15*(1), 8–26.

Christensen, P., McIntyre, N., & Pikholz, L. (2002). *Bridging community and economic development: A strategy for using industry clusters to link neighbourhoods to the regional economy*. Shorebank Enterprise Group.

Cottineau, C., & Arcaute, E. (2020). The nested structure of urban business clusters. *Applied Network Science, 5*(2). https://doi.org/10.1007/s41109-019-0246-9

Delgado, M., Porter, M. E., & Stern, S. (2014). Defining clusters of related industries (NBER Working Paper No. 20375). National Bureau of Economic Research. https://www.nber.org/papers/w20375

Dhamo, Z., Beleraj, I., & Kume, V. (2023). Business Improvement Districts: A comparative analysis of the legal framework and economic/social impact among different countries. In N. Persiani, I. E. Vannini, M. Giusti, A. Karasavvoglou, & P. Polychronidou (Eds.), *Global, regional and local perspectives on the economies of Southeastern Europe: EBEEC 2022* . Springer. https://doi.org/10.1007/978-3-031-34059-8_4

Director, H., & Raftery, A. (2022). Contour models for physical boundaries enclosing star-shaped and approximately star-shaped polygons. *Journal of the Royal Statistical Society: Series C (Applied Statistics), 71*(5), 1688–1720. https://doi.org/10.1111/rssc.12592

Ericson, W. A., (1969). Subjectives Bayesian models in sampling finite populations. Journal of the statistical society. Vol. 31, No. 2, pp. 195-233

Fisher RA (1925). "Applications of 'Student's' distribution" (PDF). *Metron*. **5**: 90–104. Archived from the original (PDF) on 5 March 2016.

Gabaix, X. (1999). Zipf's law for cities: An explanation. *The Quarterly Journal of Economics, 114*(3), 739–766.

Gabaix, X., & Ioannides, Y. (2004). The evolution of city size distributions. In J. V. Henderson & J. F. Thisse (Eds.), *Handbook of regional and urban economics* (Vol. 4, pp. 2341–2378). Elsevier.

Gabaix, X., Lasry, J., Lions, P., & Moll, B. (2016). The dynamics of inequality. *Econometrica, 84*(6), 2071–2111.

Grimaldi, R, P,. (2003). Discret and combinatorial mathematics. 5th edition.

Grodach, C., Currid-Halkett, E., Foster, N., & Murdoch III, J. (2014). The location patterns of artistic clusters: A metro- and neighborhood-level analysis. *Urban Studies, 51*(13), 2822–2843.

Hassan, A. (2012). *Stationarity and unit roots in spatial auto-regressive models* (Doctoral dissertation, Universidad Nacional de Colombia). Retrieved from https://repositorio.unal.edu.co/bitstream/handle/unal/9882/Stationarity_and_unit_roots_in_spatial_autoregressive_models.pdf?sequence=1&isAllowed=y

Hurst, Simon. "The characteristic function of the Student t distribution" . Financial Mathematics Research Report. Statistics Research Report No. SRR044-95. Archived from the original on February 18, 2010.

Li, C. H., & Calder, N. (2005). Beyond Moran's I: Testing for spatial dependence based on the spatial autoregressive model. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. https://doi.org/10.1111/j.1538-4632.2007.00708.x

Lucas, M., Sands, A., & Wolfe, D. A. (2009). Regional clusters in a global industry: ICT clusters in Canada. *European Planning Studies, 17*(2), 189–209. https://doi.org/10.1080/09654310802553415

Maoh, H., & Kanaroglou, P. (2007). Geographic clustering of firms and urban form: A multivariate analysis. *Journal of Geographical Systems, 9*(1), 29–52. https://doi.org/10.1007/s10109-006-0029-6

Meltzer, R. (2012). Understanding business improvement district formation: An analysis of neighborhoods and boundaries. *Journal of Urban Economics, 71*(1), 66–78. https://doi.org/10.1016/j.jue.2011.08.005

Moran, P. (1950). Notes on continuous stochastic phenomena. *Biometrika, 37*(1/2), 17–23.

Muller, U., & Watson, M. (2022). Spatial correlation robust inference. *Econometrica, 90*(6), 2901–2935.

Muller, U., & Watson, M. (2023). Spatial correlation robust inference in linear regression and panel models. *Journal of Business & Economic Statistics, 41*(4), 1050–1064. https://doi.org/10.1080/07350015.2022.2127737

Muller, U., & Watson, M. (2023). Spatial unit roots. Princeton University.

Niosi, J., & Bas, T. G. (2001). The competencies of regions – Canada's clusters in biotechnology. *Small Business Economics, 17*(1), 31–42. https://doi.org/10.1023/A:1011114220694

Na, J., Cao, X., Chen, J., & Chen, X. (2023). A new method for identifying the central business districts with nighttime light radiance and angular effects. *Remote Sensing, 15*(1), 239. https://doi.org/10.3390/rs15010239

OECD (2018). *Divided cities: Understanding intra-urban inequalities*. OECD Publishing. https://doi.org/10.1787/9789264300385-en

Patel, P. C. (2024). Nurturing neighborhoods, cultivating local businesses: The effects of amenities-to-infrastructure spending on new business licenses in Chicago's wards. *Journal of Business Venturing Insights, 21*, e00467.

E. Parzen. On the estimation of a probability density and the mode. Ann. Math. Statist., 33:1965–1976, 1962.

Rao, J.N.K., Scott, A.J. (1981). The analysis of categorical data from complex sample surveys: chi-squared tests for goodness of fit and independence in two-way tables. Journal of American statistical association, Vol. 76, No. 374 (Jun., 1981), pp. 221-230

M. Rosenblatt. A central limit theorem and a strong mixing condition. Proc. Nat. Acad. Sci. USA, 42:43–47, 1956.

Rozenfeld, H., Rybski, D., Gabaix, X., & Makse, H. (2011). The area and population of cities: New insights from a different perspective on cities. *American Economic Review, 101*(5), 2205–2225. http://www.aeaweb.org/articles.php?doi=10.1257/aer.101.5.2205

Sergerie, F., Chastko, K., Saunders, D., & Charbonneau, P. (2021). Defining Canada's downtown neighbourhoods: 2016 boundaries. *Statistics Canada*, Issue 2021001. Retrieved from https://www150.statcan.gc.ca/n1/pub/91f0015m/91f0015m2021001-eng.pdf

Severini, A. T. (2005). Elements of distribution theory. Cambridge University Press. https://www.cambridge.org/core/books/elements-of-distribution-theory/1091612241C8BD626046E5F1E47D30A3

Shybalkina, I. (2022). Place-based small business support and its implications for neighborhood revitalization. *Economic Development Quarterly, 36*(4), 355–370. https://doi.org/10.1177/08912424221123501

Silverman, B. W. (1986). Density estimation for statistics and data analysis. Monographs on Statistics and Applied Probability. Available online at: https://ned.ipac.caltech.edu/level5/March02/Silverman/paper.pdf.

Spencer, G. M. (2014). *Cluster atlas of Canada*. Report, Munk School of Global Affairs, University of Toronto. Available online at: https://clustercollaboration.eu/sites/default/files/international_cooperation/cluster-atlas.pdf

Spencer, G. M., Vinodrai, T., Gertler, M. S., & Wolfe, D. A. (2010). Do clusters make a difference? Defining and assessing their economic performance. *Regional Studies, 44*(6), 697–715.

Steiner, B. E., & Ali, J. (2011). Government support for the development of regional food clusters: Evidence from Alberta, Canada. *International Journal of Innovation and Regional Development, 3*(2), 186–216. https://doi.org/10.1504/IJIRD.2011.038924

Stern, M. J., & Seifert, S. C. (2010). Cultural clusters: The implications of cultural assets agglomeration for neighborhood revitalization. *Journal of Planning Education and Research, 29*(3), 262–279. https://doi.org/10.1177/0739456X09358555

Taubenböck, H., Klotz, M., Wurm, J., Schmieder, B., Wooster, M., Esch, T., & Dech, S. (2013). Delineation of central business districts in mega city regions using remotely sensed data. *Remote Sensing of Environment, 136*, 386–401. https://doi.org/10.1016/j.rse.2013.05.019

Thompson, M.E., Sedransk, J., Fang, J. and Yi, G.Y. (2022). Bayesian inference for a variance component model using pairwise composite likelihood with survey data . *Survey Methodology*, Statistics Canada, Catalogue No. 12-001-X, Vol. 48, No. 1. Paper available at http://www.statcan.gc.ca/pub/12-001-x/2022001/article/00002-eng.htm.

Toronto Region Board of Trade (2021). *Regional centres: A business district analysis*. Report. Retrieved from https://bot.com/Resources/Resource-Library/Regional-Centres-A-Business-District-Analysis

Wang, B., & Wen, B. (2021). The spatial distribution of businesses and neighborhoods: What industries match or mismatch what neighborhoods? *Habitat International, 117*, 102440.

Wheeler, C. H. (2006). Businesses don't just choose a city, they choose a specific neighborhood. *Bridges*. Federal Reserve Bank of St. Louis. Retrieved from from https://fraser.stlouisfed.org/title/bridges-federal-reserve-bank-st-louis-6272/bridges-600970

Wolfe, D. A., & Gertler, M. S. (2004). Clusters from the inside and out: Local dynamics and global linkages. *Urban Studies, 41*(5-6), 1071–1093.

Yu, W., Ai, T., & Shao, S. (2015). The analysis and delimitation of central business districts using network kernel density estimation. *Journal of Transport Geography, 45*, 32–47. https://doi.org/10.1016/j.jtrangeo.2015.04.008

Zhang, X. (2019). Building effective clusters and industrial parks. In Célestin Monga & Justin Yifu Lin (Eds.), *The Oxford handbook of structural transformation*. Oxford Handbooks. https://doi.org/10.1093/oxfordhb/9780198793847.013.13

Appendix 1: proof for the non-necessity of using probabilities of combinations for the comparative analysis of the binomial process

Acknowledging $\Pr(h)$ as probability of head, the full expression is,

$$\left(c(v,v/2)*\left(\Pr(h)\right)^{\frac{v}{2}}*\left(1-\Pr(h)\right)^{v-\frac{v}{2}}\right)\Big/\sum_{j=0}^{v}c(v,j)*\left(\Pr(h)\right)^{j}*\left(1-\Pr(h)\right)^{v-j}$$

$$>$$

$$\left(c(v,v)*\left(\Pr(h)\right)^{v}*\left(1-\Pr(h)\right)^{v-v}\right)\Big/\sum_{j=0}^{v}c(v,j)*\left(\Pr(h)\right)^{j}*\left(1-\Pr(h)\right)^{v-j}$$

The expression is then equivalent to the following, if we substitute $\Pr(h)$ with the uniformity of ½,

$$\left(c(v,v/2)*\left(\frac{1}{2}\right)^{\frac{v}{2}}*\left(1-\frac{1}{2}\right)^{v-\frac{v}{2}}\right)\Big/\sum_{j=0}^{v}c(v,j)*\left(\frac{1}{2}\right)^{j}*\left(1-\frac{1}{2}\right)^{v-j}$$

$$>$$

$$\left(c(v,v)*\left(\frac{1}{2}\right)^{v}*\left(1-\frac{1}{2}\right)^{v-v}\right)\Big/\sum_{j=0}^{v}c(v,j)*\left(\frac{1}{2}\right)^{j}*\left(1-\frac{1}{2}\right)^{v-j}$$

Simplifying with the probabilities' subtractions and exponents' subtractions, we get,

$$\left(\frac{v!}{\left(\frac{v}{2}\right)!\left(v-\frac{v}{2}\right)!}*\left(\frac{1}{2}\right)^{\frac{v}{2}}*\left(\frac{1}{2}\right)^{\frac{v}{2}}\right)\Big/\sum_{j=0}^{v}\frac{v!}{(j)!(v-j)!}*\left(\frac{1}{2}\right)^{j}*\left(\frac{1}{2}\right)^{v-j}$$

$$>$$

$$\left(\frac{v!}{(v)!(v-v)!}*\left(\frac{1}{2}\right)^{v}*\left(\frac{1}{2}\right)^{0}\right)\Big/\sum_{j=0}^{v}\frac{v!}{(j)!(v-j)!}*\left(\frac{1}{2}\right)^{j}*\left(\frac{1}{2}\right)^{v-j}$$

Now, due to the uniformity of the binomial process, we use the probability exponent's additive property to get,

$$\left(\frac{v!}{\left(\frac{v}{2}\right)!\left(v-\frac{v}{2}\right)!}*\left(\frac{1}{2}\right)^{v}\right) / \sum_{j=0}^{v}\frac{v!}{(j)!(v-j)!}*\left(\frac{1}{2}\right)^{v}$$

$$>$$

$$\left(\frac{v!}{(v)!(v-v)!}*\left(\frac{1}{2}\right)^{v}\right) / \sum_{j=0}^{v}\frac{v!}{(j)!(v-j)!}*\left(\frac{1}{2}\right)^{v}$$

finally, cancelling out common terms on both side of the inequality, and acknowledging some equalities, we get,

$$\frac{v!}{\left(\frac{v}{2}\right)!\left(v-\frac{v}{2}\right)!} > \frac{v!}{(v)!(v-v)!} = \frac{v!}{(0)!(v-0)!} = 1$$

which explains why our comparative analysis focuses on the number of combinations and not the full expression. ■

This proof makes use of a denominator equal to the sum of all probabilities, on both sides of the inequality. This denominator is technically redundant and equal to 1. However, the denominator remains useful to visualize probabilities. That is, if we cancel out all probability terms $(\Pr())$ due to the uniformity of the random process, then only combinatorial terms remain at the numerator and denominator. The denominator becomes the sum of all possible combinations and the numerator becomes the number of combinations of interest. This provides a probability in itself.

Appendix 2: Justification for applying within DB instead of within DA random allocation of jobs during pre-kernel processing

Sergerie et al, 2021, applies a uniform randomization allocation of employment within the DA boundaries. This strategy is excellent considering the number of jobs within a DA is large and covers a large set of 2-digit NAICS. However, for the case of our applications, we treat a single 2-digit NAICS at a time for the generation of the clusters. Consequently, the number of jobs involved per DA is relatively more limited. The normality approximation theory documented in this paper is conditional to a large number of jobs. That is, if the geographical surface dedicated to jobs allocation is large compared to the number of jobs itself, then reaching out to an accurate normal distribution becomes difficult. This annex explains in detail the reason for processing within DB instead of within DA random allocation of jobs during pre-kernel processing. To do so, we decompose the global random allocation of the DA with the perspective of the several DBs of the same DA. In the next pages of this annex, we are going to present equations y*, y** and y*** before presenting our final explanation.

Let's define space set $A, A' \text{ and } A''$ as the symmetric neighborhood around the centroid of the DB, the DB itself and the DA from which the DB belongs to, respectively. A is smaller than its DB and proportional in radius to the size of its DB, respectively. Let's define supplementary space set $A*'$ and $A*''$ as the superficies subtracting the previous smaller superficies. That is, $A*' = A' \setminus A$ and $A*'' = A'' \setminus A'$. Let's define S, S', S'' as the finite number of spatial spots in $A, A' \text{ and } A''$ available for random allocation of jobs. Let's also define $S*' = S' - S$ and $S*'' = S'' - S'$. Let's assume that $A < A' < A''$, which is by definition, always the case. Let's define a share of normalized probability such as $0 < \Pr(A*'')$, $0 < \Pr(A*')$, $0 < \Pr(A)$ and $\Pr(A*'') + \Pr(A*') + \Pr(A) = 1$, which define the level of smoothness or fragmentation. The 3 shares should not be proportional to the relative volume of the space sets but rather proportional to their relative importance. Let's define uniformity on $A, A*' \text{ and } A*''$. That is, $\Pr(A) * (1/S)$, $\Pr(A*') * (1/S*')$ and $\Pr(A*'') * (1/S*'')$ are the equal probabilities for all spatial spots located in $A, A*' \text{ and } A*''$, respectively. It is trivial to show that the sum of all probabilities is 1 in the sense that the space of probabilities measures is well define. That is, $(\Pr(A) * (1/S) * S) + (\Pr(A*') * (1/S*') * S*') + (\Pr(A*'') * (1/S*'') * S*'') = 1$. If the sum is 1 for the 3 $\Pr()$ terms, then the sum of the last expression remains 1, no matter the size of $S, S*' \text{ and } S*''$. An application allowing a limited amount of spread beyond the DB would be $\Pr(S) = 0.6$, $\Pr(S*') = 0.3$, and $\Pr(S*'') = 0.1$. For room convenience matter, we label the probabilities such as, $\check{Z} = \Pr(A) * (1/S)$, $\check{T} = \Pr(A*') * (1/S*')$, and $\check{N} = \Pr(A*'') * (1/S*'')$. Also, $1 - \check{Z} = \check{Z}-$, $1 - \check{T} = \check{T}-$ and $1 - \check{N} = \check{N}-$. Such a random process RP^* is not necessarily uniform anymore and still converges in distribution to the multivariate normal distribution. That is, $RP^* \to N()$, if $v \to \infty$, and in a finite context, $RP^* \approx N(vp^{*T}, vM^*)$, if v is large. v[29] is the total job count of the DB of interest (spatial set A') and vp^* is the new $(S+S*'+S*'')$-dimension expected vector, that is, $vp^* = \langle v\check{Z}, \ldots, v\check{Z}, v\check{T}, \ldots, v\check{T}, v\check{N}, \ldots, v\check{N} \rangle$, is a row vector and $M^* = P^* - p^* p^{*T}$, and $P^* = I^* p^{*T}$, and I^* is the identity matrix of dimension $(S+S*'+S*'')$ x $(S+S*'+S*'')$. That is, P^* is a diagonal matrix whose diagonal elements are the items of vector p^*. In other words, $P^* = I^* p^{*T} =$

29. v doesn't use the '*' notation because it is the same v as the one presented in the main methodology of this paper. That is, the number of v unique job of within a DB.

$$\begin{bmatrix}
\check{Z} & 0 & \dots & 0 & 0 & 0 & \dots & 0 & 0 & 0 & \dots & 0 \\
0 & \check{Z} & \dots & 0 & 0 & 0 & \dots & 0 & 0 & 0 & \dots & 0 \\
\vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\
0 & 0 & \dots & \check{Z} & 0 & 0 & \dots & 0 & 0 & 0 & \dots & 0 \\
0 & 0 & \dots & 0 & \check{T} & 0 & \dots & 0 & 0 & 0 & \dots & 0 \\
0 & 0 & \dots & 0 & 0 & \check{T} & \dots & 0 & 0 & 0 & \dots & 0 \\
\vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\
0 & 0 & \dots & 0 & 0 & 0 & \dots & \check{T} & 0 & 0 & \dots & 0 \\
0 & 0 & \dots & 0 & 0 & 0 & \dots & 0 & \check{N} & 0 & \dots & 0 \\
0 & 0 & \dots & 0 & 0 & 0 & \dots & 0 & 0 & \check{N} & \dots & 0 \\
\vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\
0 & 0 & \dots & 0 & 0 & 0 & \dots & 0 & 0 & 0 & \dots & \check{N}
\end{bmatrix},$$

$$\text{And } p^{*}p^{*T} = \check{Z}\check{Z} + \dots + \check{Z}\check{Z} + \check{T}\check{T} + \dots + \check{T}\check{T} + \check{N}\check{N} + \dots + \check{N}\check{N}$$

For its part, vM^{*} is the new variance-covariance matrix, that is, the full uncertainty quantification can be represented as vM^{*} =

$$\begin{bmatrix}
v\check{Z}(\check{Z}-) & -v\check{Z}\check{Z} & \cdots & -v\check{Z}\check{Z} & -v\check{Z}\check{T} & -v\check{Z}\check{T} & \cdots & -v\check{Z}\check{T} & -v\check{Z}\check{N} & -v\check{Z}\check{N} & \cdots & -v\check{Z}\check{N} \\
-v\check{Z}\check{Z} & v\check{Z}(\check{Z}-) & \cdots & -v\check{Z}\check{Z} & -v\check{Z}\check{T} & -v\check{Z}\check{T} & \cdots & -v\check{Z}\check{T} & -v\check{Z}\check{N} & -v\check{Z}\check{N} & \cdots & -v\check{Z}\check{N} \\
\vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\
-v\check{Z}\check{Z} & -v\check{Z}\check{Z} & \cdots & v\check{Z}(\check{Z}-) & -v\check{Z}\check{T} & -v\check{Z}\check{T} & \cdots & -v\check{Z}\check{T} & -v\check{Z}\check{N} & -v\check{Z}\check{N} & \cdots & -v\check{Z}\check{N} \\
. & . & \cdots & . & v\check{T}(\check{T}-) & -v\check{T}\check{T} & \cdots & -v\check{T}\check{T} & -v\check{T}\check{N} & -v\check{T}\check{N} & \cdots & -v\check{T}\check{N} \\
. & . & \cdots & . & -v\check{T}\check{T} & v\check{T}(\check{T}-) & \cdots & -v\check{T}\check{T} & -v\check{T}\check{N} & -v\check{T}\check{N} & \cdots & -v\check{T}\check{N} \\
\vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\
. & . & \cdots & . & -v\check{T}\check{T} & -v\check{T}\check{T} & \cdots & v\check{T}(\check{T}-) & -v\check{T}\check{N} & -v\check{T}\check{N} & \cdots & -v\check{T}\check{N} \\
. & . & \cdots & . & . & . & \cdots & . & v\check{N}(\check{N}-) & -v\check{N}\check{N} & \cdots & -v\check{N}\check{N} \\
. & . & \cdots & . & . & . & \cdots & . & -v\check{N}\check{N} & v\check{N}(\check{N}-) & \cdots & -v\check{N}\check{N} \\
\vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\
. & . & \cdots & . & . & . & \cdots & . & -v\check{N}\check{N} & -v\check{N}\check{N} & \cdots & v\check{N}(\check{N}-)
\end{bmatrix}$$

vM^{*} is a finite block matrix whose block diagonal includes 3 sub-matrices, one for A, $A*'$, and $A*''$, respectively. The dimensions of the 3 sub-matrices are $S \times S$, $S*' \times S*'$, and $S*'' \times S*''$, respectively. Consequently, variance-covariance matrix vM^{*} is symmetric and has dimension $(S+S*'+S*'') \times (S+S*'+S*'')$. vM^{*} is also a generalization of vM, presented in this paper. That is, vM^{*} reduces to vM and is equal to the first two of the 3 sub-matrices in the block diagonal of vM^{*}, if $Pr(A*'')$ is non existent, and $Pr(A) + Pr(A*') = 1$, and $Pr(A) * (1/S) = Pr(A*') * (1/S*')$. Matrix terms below the block diagonal are omitted for clarity matter and also to make the 3 sub-matrices of the block diagonal more obvious. A matrix term below the block diagonal is equal to the term of the inverse coordinate in the above block diagonal. vM^{*} is slightly more sophisticated than vM due to the 3-level spatial structure involving A, $A*'$, and $A*''$. The new multivariate normal distribution of our converging multinomial random process can be expressed[30] in a finite context in the following manner,

30. All vectors and matrices dimensions are technically based on $(S+S*'+S*'')-1$, and not $(S+S*'+S*'')$. We keep $(S+S*'+S*'')$ for simplification matter in the notation. By definition, the last category of a multinomial random variable is a redundant information because the sum of counts per categories is a fixed total constraint, and the count of jobs allocated to the last category is automatically deducted by the sum of all other categories. Consequently, the variance-covariance matrix is rank deficient and not full rank. To avoid singularity, the impossibility of matrix inversion, and improper matrix determinant, the dimension of the multivariate normal distribution needs to collapse within a sub-space of $(S+S*'+S*'')$, which is simply, $(S+S*'+S*'')-1$.

$$y^{*} = \mathrm{MVN}\left(x^{*}; vp^{*}, vM^{*}\right) = \left(2\pi\right)^{-\frac{S+S^{**}+S^{***}}{2}} * \det\left(vM^{*}\right)^{-\frac{1}{2}} * \exp\left(-\frac{1}{2}\left(x^{*}-vp^{*}\right)\left(vM^{*}\right)^{-1}\left(x^{*}-vp^{*}\right)^{T}\right) \text{and} \, x^{*} \approx \mathrm{N}\left(vp^{*T}, vM^{*}\right)$$

What is more? Let's define ADA as the set of All DB of the dissemination Area. The material presented so far in this annex is about the random allocation of a population of v unique jobs of a DB in its ADA. However, the same should be done for all other DBs of the same ADA because each DB of the ADA will randomly allocate its own v jobs across the DBs of the same ADA. Consequently, we now must consider the joint multivariate normal distribution instead of a single multivariate normal distribution. Fundamentally, it is about a simple concatenation exercise of the material presented previously in this annex. To do so, it is sufficient to keep the existing notation of this annex and simply introduce the DB index and make it range from $\mathrm{DB} = 1,2,...,q(\mathrm{ADA})$, where $q(\mathrm{ADA})$ is the number of total unique DB in the ADA where jobs data points are available. From previous material in this annex, we know that each random process, one for each DB, converges in distribution to the multivariate normal distribution. That is,

$$\mathrm{RP}(1)^{*} \to \mathrm{N}(), \text{ if } \mathrm{v}(1) \to \infty,$$

$$\mathrm{RP}(2)^{*} \to \mathrm{N}(), \text{ if } \mathrm{v}(2) \to \infty,$$

...,

$$\mathrm{RP}\left(\mathrm{q}(\mathrm{ADA})\right)^{*} \to \mathrm{N}(), \text{ if } \mathrm{v}\left(\mathrm{q}(\mathrm{ADA})\right) \to \infty$$

In a finite context, this is equivalent to say that,

$$\mathrm{RP}(1)^{*} \approx \mathrm{N}\left(\mathrm{v}(1)\mathrm{p}(1)^{*T}, \mathrm{v}(1)\mathrm{M}(1)^{*}\right), \text{ if } \mathrm{v}(1) \text{ is large},$$

$$\mathrm{RP}(2)^{*} \approx \mathrm{N}\left(\mathrm{v}(2)\mathrm{p}(2)^{*T}, \mathrm{v}(2)\mathrm{M}(2)^{*}\right), \text{ if } \mathrm{v}(2) \text{ is large},$$

...,

$$\mathrm{RP}\left(\mathrm{q}(\mathrm{ADA})\right)^{*} \approx \mathrm{N}\left(\mathrm{v}\left(\mathrm{q}(\mathrm{ADA})\right)\mathrm{p}\left(\mathrm{q}(\mathrm{ADA})\right)^{*T}, \mathrm{v}\left(\mathrm{q}(\mathrm{ADA})\right)\mathrm{M}\left(\mathrm{q}(\mathrm{ADA})\right)^{*}\right), \text{ if } \mathrm{v}\left(\mathrm{q}(\mathrm{ADA})\right) \text{ is large}$$

It is also equivalent to say that,

$$\mathrm{x}(1)^{*} \approx \mathrm{N}\left(\mathrm{v}(1)\mathrm{p}(1)^{*T}, \mathrm{v}(1)\mathrm{M}(1)^{*}\right), \text{ if } \mathrm{v}(1) \text{ is large},$$

$$\mathrm{x}(2)^{*} \approx \mathrm{N}\left(\mathrm{v}(2)\mathrm{p}(2)^{*T}, \mathrm{v}(2)\mathrm{M}(2)^{*}\right), \text{ if } \mathrm{v}(2) \text{ is large},$$

...,

$$\mathrm{x}\left(\mathrm{q}(\mathrm{ADA})\right)^{*} \approx \mathrm{N}\left(\mathrm{v}\left(\mathrm{q}(\mathrm{ADA})\right)\mathrm{p}\left(\mathrm{q}(\mathrm{ADA})\right)^{*T}, \mathrm{v}\left(\mathrm{q}(\mathrm{ADA})\right)\mathrm{M}\left(\mathrm{q}(\mathrm{ADA})\right)^{*}\right), \text{ if } \mathrm{v}\left(\mathrm{q}(\mathrm{ADA})\right) \text{ is large}$$

Each of the multivariate normal distribution above refer to a distinct space of probabilities measures. That is, the probability vector and the set of employees (jobs) is distinct. However, the set of spatial spots is the same. Now, making use of the product operator, and assuming the independence of the several distributions, the multivariate normal distributions can be expressed in a single equation and generate the joint multivariate normal distribution y^{**}, that is,

$$y^{**} = y(1)^{*} * y(2)^{*} * \ldots * y\left(q(\mathrm{ADA})\right)^{*}$$

$$= \mathrm{JMVN}\left(x(1)^{*}, x(2)^{*}, \ldots, x\left(q(\mathrm{ADA})\right)^{*}; v(1)p(1)^{*}, v(2)p(2)^{*}, \ldots, v\left(q(\mathrm{ADA})\right)p\left(q(\mathrm{ADA})\right)^{*}, \right.$$
$$\left. v(1)M(1)^{*}, v(2)M(2)^{*}, \ldots, v\left(q(\mathrm{ADA})\right)M\left(q(\mathrm{ADA})\right)^{*} \right)$$

$$= (2\pi)^{-\frac{S+S^{*'}+S^{*''}}{2}} * \det\left(v(1)M(1)^{*}\right)^{-\frac{1}{2}} * \exp\left(-\frac{1}{2}\left(x(1)^{*} - v(1)p(1)^{*}\right) * \left(v(1)M(1)^{*}\right)^{-1} * \left(x(1)^{*} - v(1)p(1)^{*}\right)^{\mathrm{T}}\right)$$

$$* (2\pi)^{-\frac{S+S^{*'}+S^{*''}}{2}} * \det\left(v(2)M(2)^{*}\right)^{-\frac{1}{2}} * \exp\left(-\frac{1}{2}\left(x(2)^{*} - v(2)p(2)^{*}\right) * \left(v(2)M(2)^{*}\right)^{-1} * \left(x(2)^{*} - v(2)p(2)^{*}\right)^{\mathrm{T}}\right)$$

$$* \ldots *$$

$$(2\pi)^{-\frac{S+S^{*'}+S^{*''}}{2}} * \det\left(v\left(q(\mathrm{ADA})\right)M\left(q(\mathrm{ADA})\right)^{*}\right)^{-\frac{1}{2}}$$
$$* \exp\left(-\frac{1}{2}\left(x\left(q(\mathrm{ADA})\right)^{*} - v\left(q(\mathrm{ADA})\right)p\left(q(\mathrm{ADA})\right)^{*}\right) * \left(v\left(q(\mathrm{ADA})\right)M\left(q(\mathrm{ADA})\right)^{*}\right)^{-1} * \left(x\left(q(\mathrm{ADA})\right)^{*} - v\left(q(\mathrm{ADA})\right)p\left(q(\mathrm{ADA})\right)^{*}\right)^{\mathrm{T}}\right)$$

$$= (2\pi)^{-\frac{q(\mathrm{ADA}) * \left(S+S^{*'}+S^{*''}\right)}{2}} * \left(\prod_{\mathrm{DB}=1}^{\mathrm{DB}=q(\mathrm{ADA})} \det\left(v(\mathrm{DB})M(\mathrm{DB})^{*}\right)^{-\frac{1}{2}}\right)$$
$$* \left(\exp\left(\sum_{\mathrm{DB}=1}^{\mathrm{DB}=q(\mathrm{ADA})} -\frac{1}{2}\left(x(\mathrm{DB})^{*} - v(\mathrm{DB})p(\mathrm{DB})^{*}\right) * \left(v(\mathrm{DB})M(\mathrm{DB})^{*}\right)^{-1} * \left(x(\mathrm{DB})^{*} - v(\mathrm{DB})p(\mathrm{DB})^{*}\right)^{\mathrm{T}}\right)\right)$$

Looking at the third equality, we can notice that the term $(S + S^{*'} + S^{*''})$ doesn't make use of the DB index. That is, it is used repetitively for each multivariate normal equation because the value doesn't change across DB. Looking at the fourth equality (the last equality), the summation is justified from the exponent property under the common term $\exp$ and is used to diminish the length of the full expression. It is worth noting that, the joint distribution y^{**} reaches its maximum density if each vector $x(1)^{*}, x(2)^{*}, \ldots, x\left(q(\mathrm{ADA})\right)^{*}$ is equal to their mean vector $v(1)p(1)^{*}, v(2)p(2)^{*}, \ldots, v\left(q(\mathrm{ADA})\right)p\left(q(\mathrm{ADA})\right)^{*}$, respectively. The same way around, y^{**} reaches its minimum density if each vector $x(1)^{*}, x(2)^{*}, \ldots, x\left(q(\mathrm{ADA})\right)^{*}$ allocates the totality of their resources to the element of the probability vector $p(1)^{*}, p(2)^{*}, \ldots, p\left(q(\mathrm{ADA})\right)^{*}$ where the probability is at its lowest possible, respectively. Furthermore, each input vector $x(1)^{*}, x(2)^{*}, \ldots, x\left(q(\mathrm{ADA})\right)^{*}$ of the joint distribution y^{**} is limited to its total number of available jobs $v(1), v(2), \ldots, v\left(q(\mathrm{ADA})\right)$, respectively. Consequently, the multivariate normal distributions $y(1)^{*}, y(2)^{*}, \ldots, y\left(q(\mathrm{ADA})\right)^{*}$ are fully independent of each other, in the sense that, the random allocation of jobs of one DB within the ADA doesn't provide any information about the way another DB will allocate its own resources in the same ADA. In other words, we can write the following variance-covariance matrix M^{**} to summarize the setting. In other words,

$$M^{**} = \begin{bmatrix} v(1)M(1)^{*} & 0 & \cdots & 0 \\ 0 & v(2)M(2)^{*} & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & v\left(q(ADA)\right)M\left(q(ADA)\right)^{*} \end{bmatrix}$$

Where, M^{**} is the variance-covariance matrix of the joint multivariate normal of vector of vectors $\left(x(1)^{*}, x(2)^{*}, \ldots, x\left(q(ADA)\right)^{*}\right)$, that is,

$$x^{**} = \left(x(1)^{*}, x(2)^{*}, \ldots, x\left(q(ADA)\right)^{*}\right) \approx N\left(\left(v(1)p(1)^{*T}, v(2)p(2)^{*T}, \ldots, v\left(q(ADA)\right)p\left(q(ADA)\right)^{*T}\right), M^{**}\right), \text{ if}$$
$$v(1), v(2), \ldots, v\left(q(ADA)\right) \text{ is large, respectively.}$$

Each element of the matrix M^{**} has been introduced previously. M^{**} is only a way to re-package existing information to make ideas clearer. The dimension of M^{**} is $q(ADA)(S + S*' + S*'') * q(ADA)(S + S*' + S*'')$ [31]. M^{**} must be seen as a nested block matrix. That is, M^{**} is a diagonal block matrix, where each element on the diagonal is itself a block matrix of dimension $(S + S*' + S*'') \text{ x } (S + S*' + S*'')$ representing inter-connections within a single DB, and each element off-diagonal is a block matrix of dimension $(S + S*' + S*'') \text{ x } (S + S*' + S*'')$ populated entirely by zeros representing the absence of inter-connection between a pair of distinct DB part of the same ADA and with available data. The latter is nothing more than the matrix representation of the full independence of the several multivariate normal distributions justifying the use of the product operator in the previous joint multivariate normal equation.

The material presented over the previous page of this annex propose a joint distribution of $q(ADA)$ distinct multivariate normal where each one is dedicated to a specific DB of the same ADA. The input variable is a vector of vectors $\left(x(1)^{*}, x(2)^{*}, \ldots, x\left(q(ADA)\right)^{*}\right)$ and require the analyst to input $q(ADA) * (S + S*' + S*'')$ distinct value. This number of inputs to provide is large and a burden for the analyst. For this reason, we re-package the same original ''joint'' idea into an alternative model where the input variable is of dimension $(S + S*' + S*'')$ only. This alternative model make use of a single multivariate normal, however the mean vector is a sum of all $q(ADA)$ mean vectors and the variance-covariance matrix is a sum of all $q(ADA)$ variance-covariance matrices of the $q(ADA)$ multivariate normal distributions. Formally, in a finite context, we have,

$$x^{***} = x(1)^{*} + x(2)^{*} + \ldots + x\left(q(ADA)\right)^{*} \approx N\left(p^{**T}, M^{***}\right),$$
$$\text{If } v(1), v(2), \ldots, v\left(q(ADA)\right) \text{ is large, respectively.}$$

Where, x^{***} is a row vector of dimension $(S + S*' + S*'')$, where the sum of terms of x^{***} is the scalar VDB, that is,

$$x^{***}\begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} = \left(1 * v(1)\right) + \left(1 * v(2)\right) + \ldots + \left(1 * v\left(q(ADA)\right)\right) = VDB$$

31. All vectors and matrices dimensions are technically based on $(S+S*'+S*'')-1$, and not $(S+S*'+S*'')$. We keep $(S+S*'+S*'')$ for simplification matter in the notation. By definition, the last category of a multinomial random variable is a redundant information because the sum of counts per categories is a fixed total constraint, and the count of jobs allocated to the last category is automatically deducted by the sum of all other categories. Consequently, the variance-covariance matrix is rank deficient and not full rank. To avoid singularity, the impossibility of matrix inversion, and improper matrix determinant, the dimension of the multivariate normal distribution needs to collapse within a sub-space of $(S+S*'+S*'')$, which is simply, $(S+S*'+S*'')-1$. In the specific case of M^{**}, the dimension is based on $q(ADA)((S+S*'+S*'')-1)$, and not $q(ADA)(S+S*'+S*'')$, because each of the $q(ADA)$ matrices need the adjustment within the block matrix M^{**}. Same thing for the input vector and the expected vector of the normal joint multivariate distribution.

And where,

$$E\left(x^{***}\right) = p^{**} = \sum_{DB=1}^{DB=q(ADA)} v\left(DB\right)p\left(DB\right)^{*}$$

And where,

$$var\left(x^{***}\right) = M^{***} = \sum_{DB=1}^{DB=q(ADA)} v\left(DB\right)M\left(DB\right)^{*}$$

Under independence of DB level variables, and where the new multivariate normal distribution can be expressed as[32],

$$y^{***} = MVN\left(x^{***}; p^{**}, M^{***}\right)$$
$$= \left(2\pi\right)^{\frac{S+S^{*'}+S^{*''}}{2}} * det\left(M^{***}\right)^{-\frac{1}{2}} * exp\left(-\frac{1}{2}\left(x^{***} - p^{**}\right) * \left(M^{***}\right)^{-1} * \left(x^{***} - p^{**}\right)^{T}\right)$$

y^{***} differ significantly from y^{**} presented earlier in this annex. The input variable x^{***} specify the allocation of VDB unique jobs within a single input vector of dimension $(S + S*' + S*'')$, instead of $q(ADA)$ distinct vectors of dimension $(S + S*' + S*'')$ each. x^{***} is approximated by a multivariate normal distribution since the exact distribution would be a convolution of several multinomials not available in closed form.

Finally, we can now describe the randomization process of Sergerie et al, 2021 as a special case, where for all DB of a same DA with data available, we have,

$$Pr\left(A\left(DB\right)\right) > 0,\ Pr\left(A\left(DB\right)*'\right) > 0,\ Pr\left(A\left(DB\right)*''\right) > 0$$

$$Pr\left(A\left(DB\right)\right) + Pr\left(A\left(DB\right)*'\right) + Pr\left(A\left(DB\right)*''\right) = 1, \text{ and}$$

$$Pr\left(A\left(DB\right)\right) * \left(1/S\left(DB\right)\right) = Pr\left(A\left(DB\right)*'\right) * \left(1/S\left(DB\right)*'\right) = Pr\left(A\left(DB\right)*''\right) * \left(1/S\left(DB\right)*''\right).$$

The last condition guarantees that all unique jobs of the DA are randomly and uniformly allocated within the DA. Also, the main randomization method of this paper is a special case if,

$$Pr\left(A*''\right) > 0 \text{ is non existent, } Pr\left(A\left(DB\right)\right) > 0 \quad Pr\left(A\left(DB\right)*'\right) > 0$$

$$Pr\left(A\left(DB\right)\right) + Pr\left(A\left(DB\right)*'\right) = 1 \text{ and}$$

$$Pr\left(A\left(DB\right)\right) * \left(1/S\left(DB\right)\right) = Pr\left(A\left(DB\right)*'\right) * \left(1/S\left(DB\right)*'\right),$$

for all DB of the ADA with available data. (doesn't apply to y***)

32. All vectors and matrices dimensions are technically based on $(S+S*'+S*'')-1$, and not $(S+S*'+S*'')$. We keep $(S+S*'+S*'')$ for simplification matter in the notation. By definition, the last category of a multinomial random variable is a redundant information because the sum of counts per categories is a fixed total constraint, and the count of jobs allocated to the last category is automatically deducted by the sum of all other categories. Consequently, the variance-covariance matrix is rank deficient and not full rank. To avoid singularity, the impossibility of matrix inversion, and improper matrix determinant, the dimension of the multivariate normal distribution needs to collapse within a sub-space of $(S+S*'+S*'')$, which is simply, $(S+S*'+S*'')-1$. In the specific case of matrix M^{***}, the sum of $q(ADA)$ matrices of dimension $(S+S*'+S*'')-1$ is also of dimension $(S+S*'+S*'')-1$.

Now that y*, y** and y*** have been introduced in detail, we can present our justification of why a within DB random allocation of jobs is preferred for the practical case of our application.

With y**, under a large geospatial surface such as a DA, there is 2 items to assume; 1) each multivariate normal distribution $\mathrm{y}(1)^{*},\mathrm{y}(2)^{*},\ldots,\mathrm{y}(\mathrm{q}(\mathrm{ADA}))^{*}$ is assumed to be well specified as a normal distribution, respectively. More precisely, for each multivariate distribution, the expected rate of occurrence λ of the polygonal space interval of interest, namely, a single spatial spot or a single pixel in the DA, is more than moderate. That is, expected outcomes vectors $\mathrm{v}(1)\mathrm{p}(1)^{*},\mathrm{v}(2)\mathrm{p}(2)^{*},\ldots,\mathrm{v}(\mathrm{q}(\mathrm{ADA}))\mathrm{p}(\mathrm{q}(\mathrm{ADA}))^{*}$ are respectively made of $(\mathrm{S}+\mathrm{S}^{*\prime}+\mathrm{S}^{*\prime\prime})$ distinct elements equal or greater than 10. In other words, we assume here that a large enough number of jobs v is present in each DB of the ADA where data are available and the number of spatial spots available in the DA is not large enough to lower down the ratio of the average number of jobs per spatial spot, below 10. Otherwise, a multivariate Poisson distribution (a multivariate right skewed normal getting squeezed toward its multivariate origin coordinate) would be more suitable for some of the distributions composing the joint distribution and the shape of the joint normal could be altered. 2) On the same topic, some probability vector p could potentially be made of probability terms close to zero due to the large number of categories. This could challenge the good approximation toward the normal distribution. However, we count on the availability of a large v per DB where data is available to compensate for the issue of sparcity and bring back the required quality of normal approximation. The intuition here, is similar to a dataset made of several variable, where the number of observations is large enough to handle the many dimensions of information.

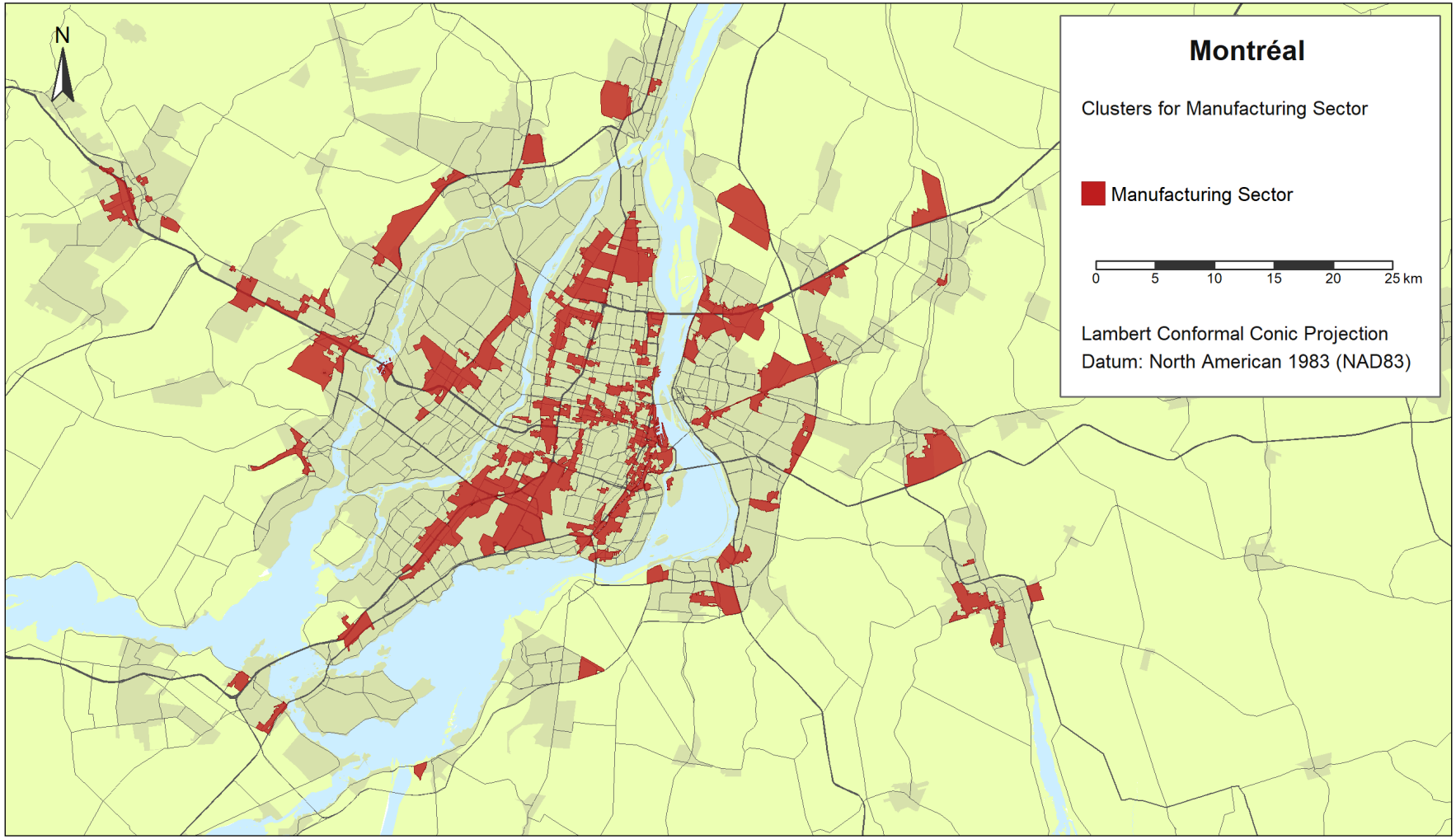
Now, for the cases of our own application, if each DB of the ADA with available data allocates jobs data points from a single 2-digit NAICS within the DA, then it is unlikely that a minimum of 10 jobs will succeed to cover each spatial spot of the DA. Therefore, it is unlikely to produce accurate marginal normal approximations. It is also likely that the accuracy of the joint normal will be altered, even under uniformity of probabilities for the full DA. Consequently, the expected outcome won't be a uniform allocation of jobs within the DA. It will be a sparce allocation not stable over random realizations. The property of smoothness will be lost. A larger kernel bandwidth will be needed to compensate for the instability and scarcity, and local accuracy of clusters will be lost. For this first reason, a random allocation within DB seems to be preferable.

A solution could be y***. With y***, it becomes more likely that a category will get a larger number of jobs randomly allocated because it is about the sum of counts from the DBs of the ADA with available data. However, for the cases of our own application, it is still unlikely that a single 2-digit NAICS will succeed to reach 10 jobs per spatial spot. For this second reason, a random allocation within DB seems to be preferable again.

A final alternative solution could be to aggregate uniformly the spatial spots of the DA and reduce the number of categories in such a way that the total number of jobs available in the DA divided per the number of aggregated categories is now equal to 10. However, for the cases of our own application, under a single 2-digit NAICS, the number of aggregated categories is likely to be not so large. A normal approximation will be reliable, and the expected outcome will be a uniform and stable allocation of jobs within the DA. However, again, a larger kernel bandwidth will be needed to compensate for the lack of high resolution and local accuracy of clusters will be lost. For this third reason, a random allocation within DB seems to be preferable again. An accurate normal approximation is not always easy to reach out to. However, in the case of our applications, a random allocation within DB remains our best option. Documenting the normal approximation in this paper remain essential, because if you have a deviation, the best way to understand this deviation is to understand what you are deviating from in the first place. ■

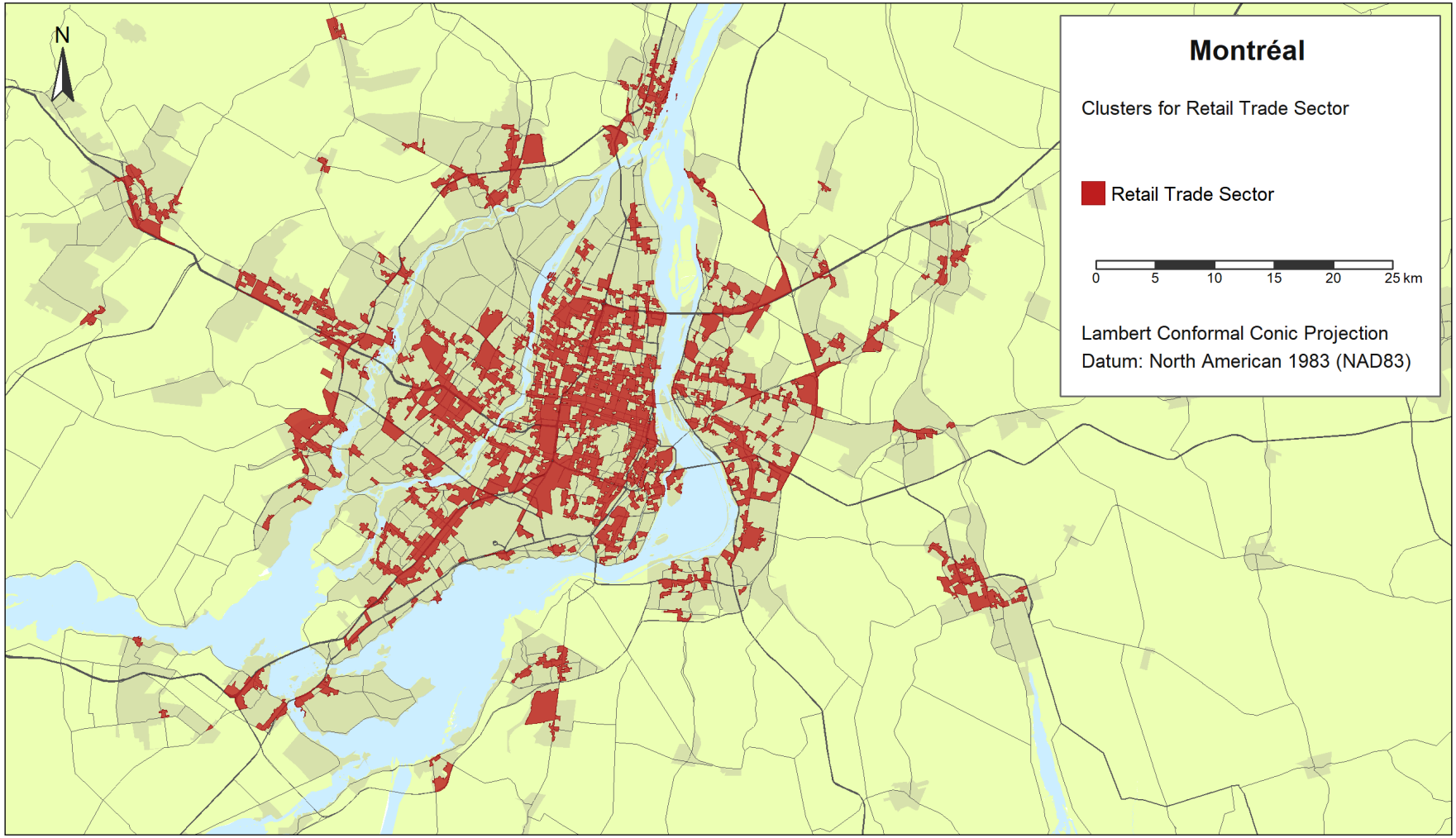
Appendix 3: cluster mapping results

Map 5
Montréal Manufacturing Sector



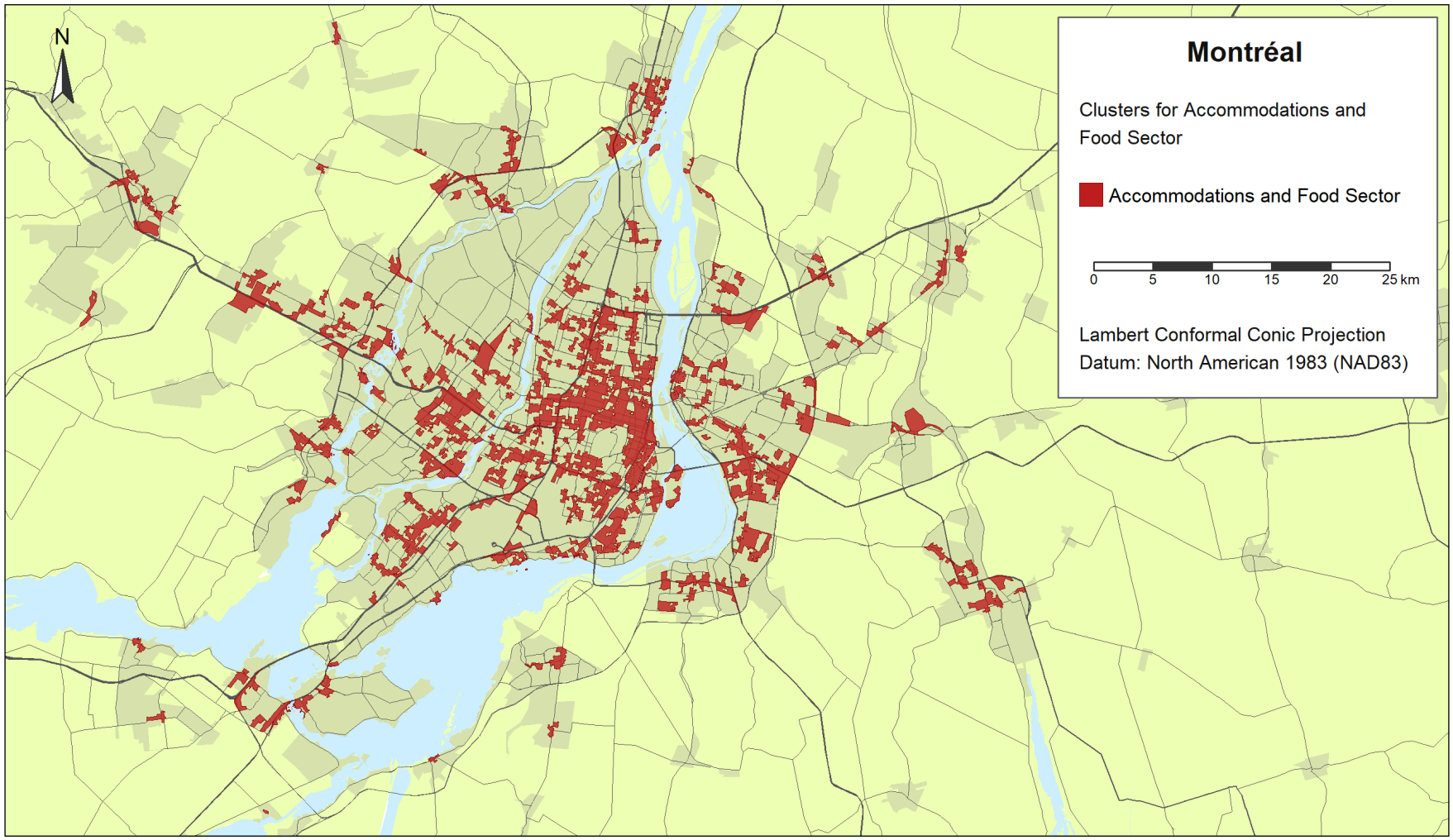
Note: 26% of DBs within the CMA, containing 90% of the employees in this industry sector
Sources: Statistics Canada, 2021 Census – Provinces/territories boundary file, 2021 Census – Population centres boundary file, 2021 Census – Road network file, 2016 Census - Coastal waters boundary file, 2016 Census - Lakes and Rivers boundary file, and authors' computations from the BR database.

Map 6
Montréal Retail Trade Sector



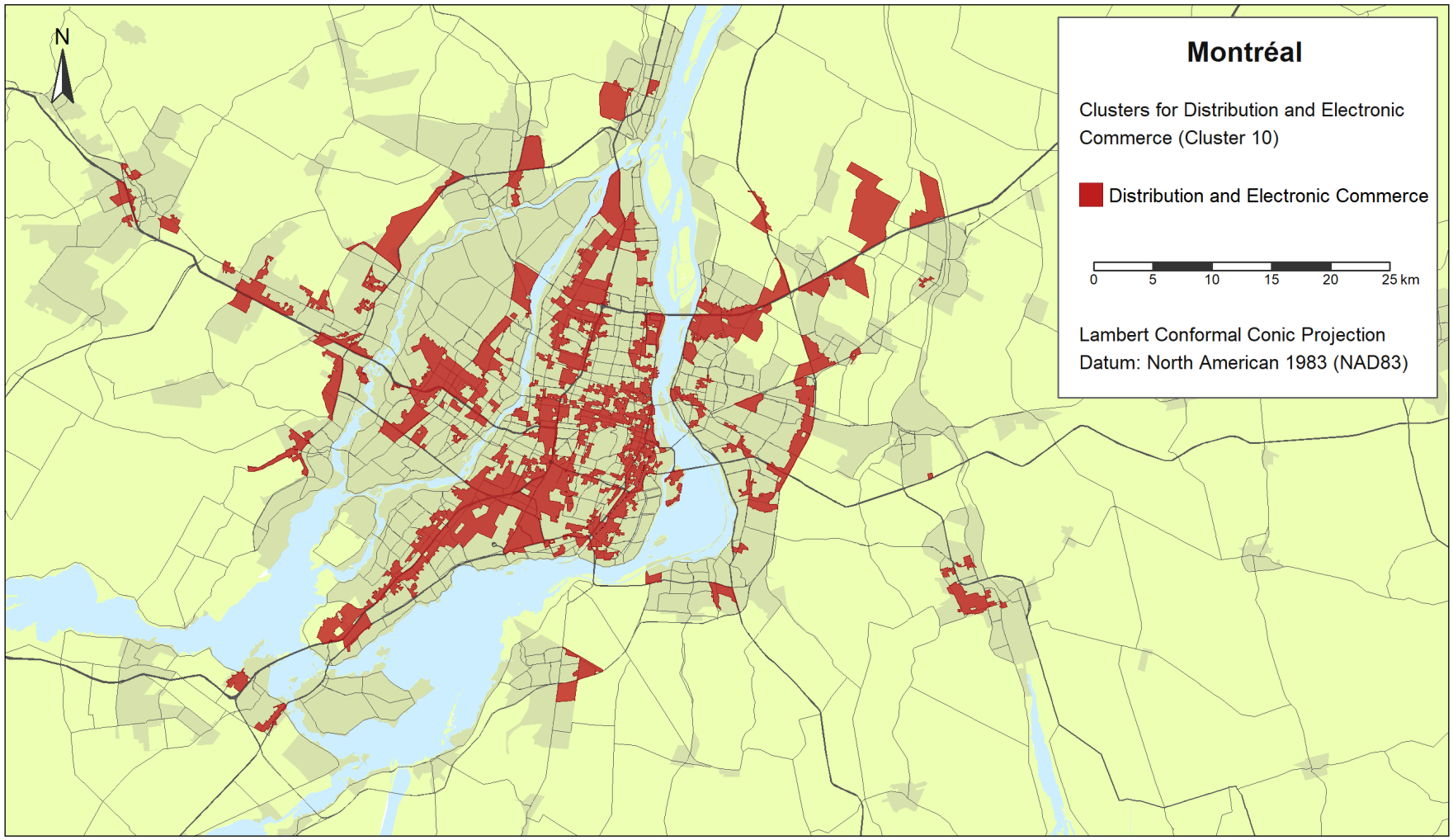
Note: 38% of DBs within the CMA, containing 92% of the employees in this industry sector
Sources: Statistics Canada, 2021 Census – Provinces/territories boundary file, 2021 Census – Population centres boundary file, 2021 Census – Road network file, 2016 Census - Coastal waters boundary file, 2016 Census - Lakes and Rivers boundary file, and authors' computations from the BR database.

Map 7
Montréal Accommodations and Food Services Sector



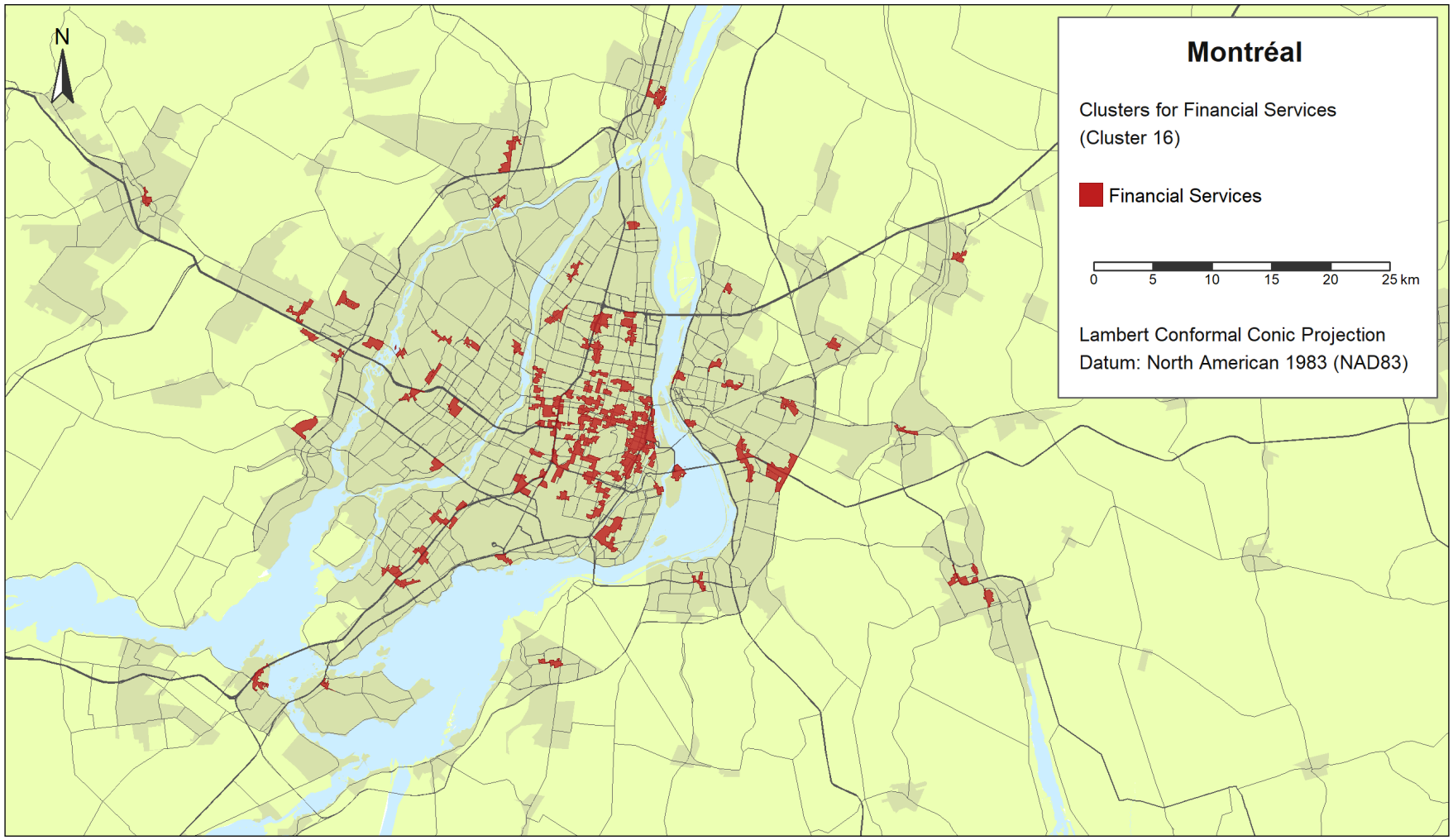
Note: 36% of DBs within the CMA, containing 85% of the employees in this industry sector
Sources: Statistics Canada, 2021 Census – Provinces/territories boundary file, 2021 Census – Population centres boundary file, 2021 Census – Road network file, 2016 Census - Coastal waters boundary file, 2016 Census - Lakes and Rivers boundary file, and authors' computations from the BR database.

Map 8
Montréal Distribution and Electronic Commerce (cluster 10)



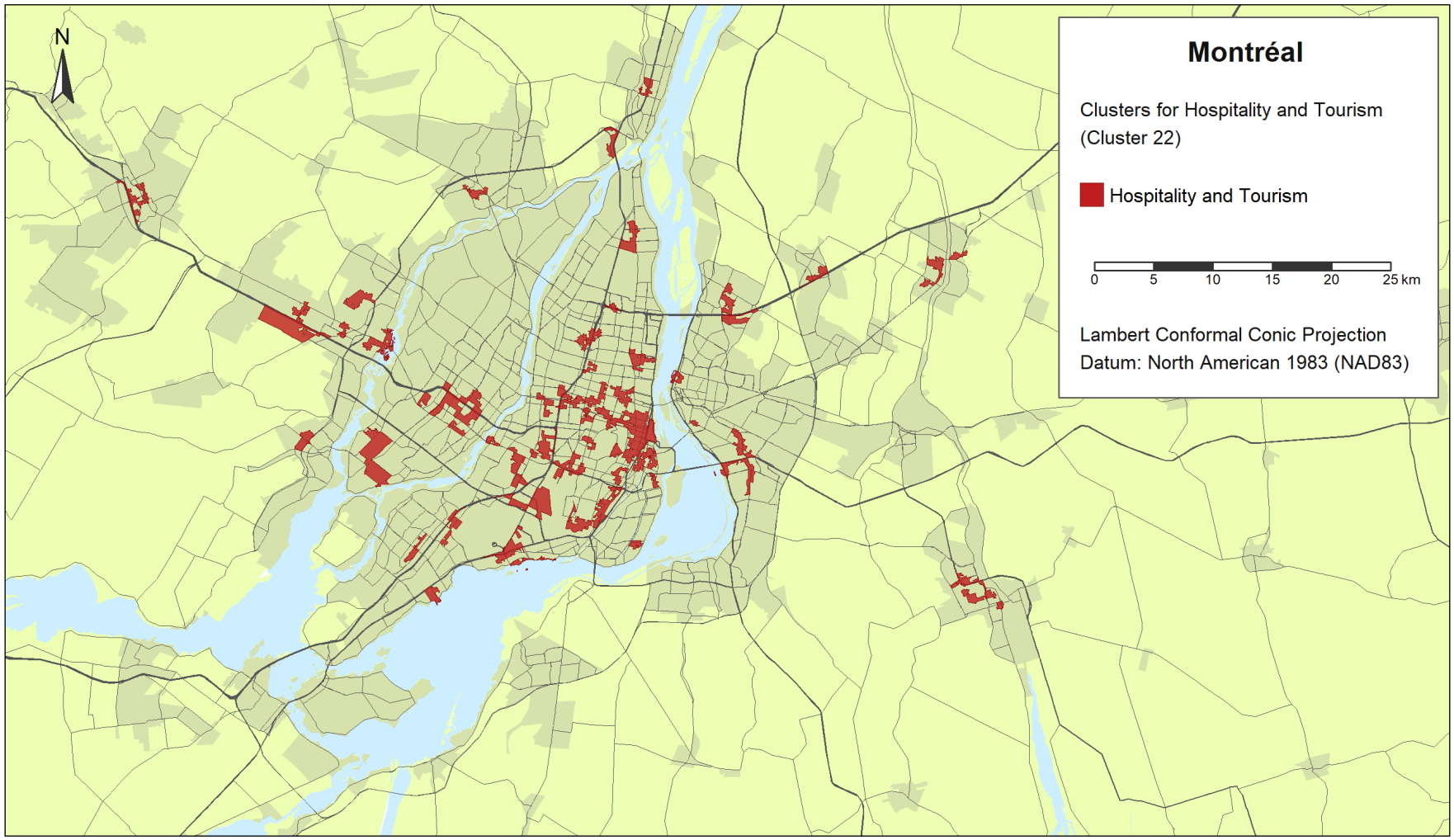
Note: 34% of DBs within the CMA, containing 95% of the employees in this industry sector
Sources: Statistics Canada, 2021 Census – Provinces/territories boundary file, 2021 Census – Population centres boundary file, 2021 Census – Road network file, 2016 Census - Coastal waters boundary file, 2016 Census - Lakes and Rivers boundary file, and authors' computations from the BR database.

Map 9
Montréal Financial Services (cluster 16)



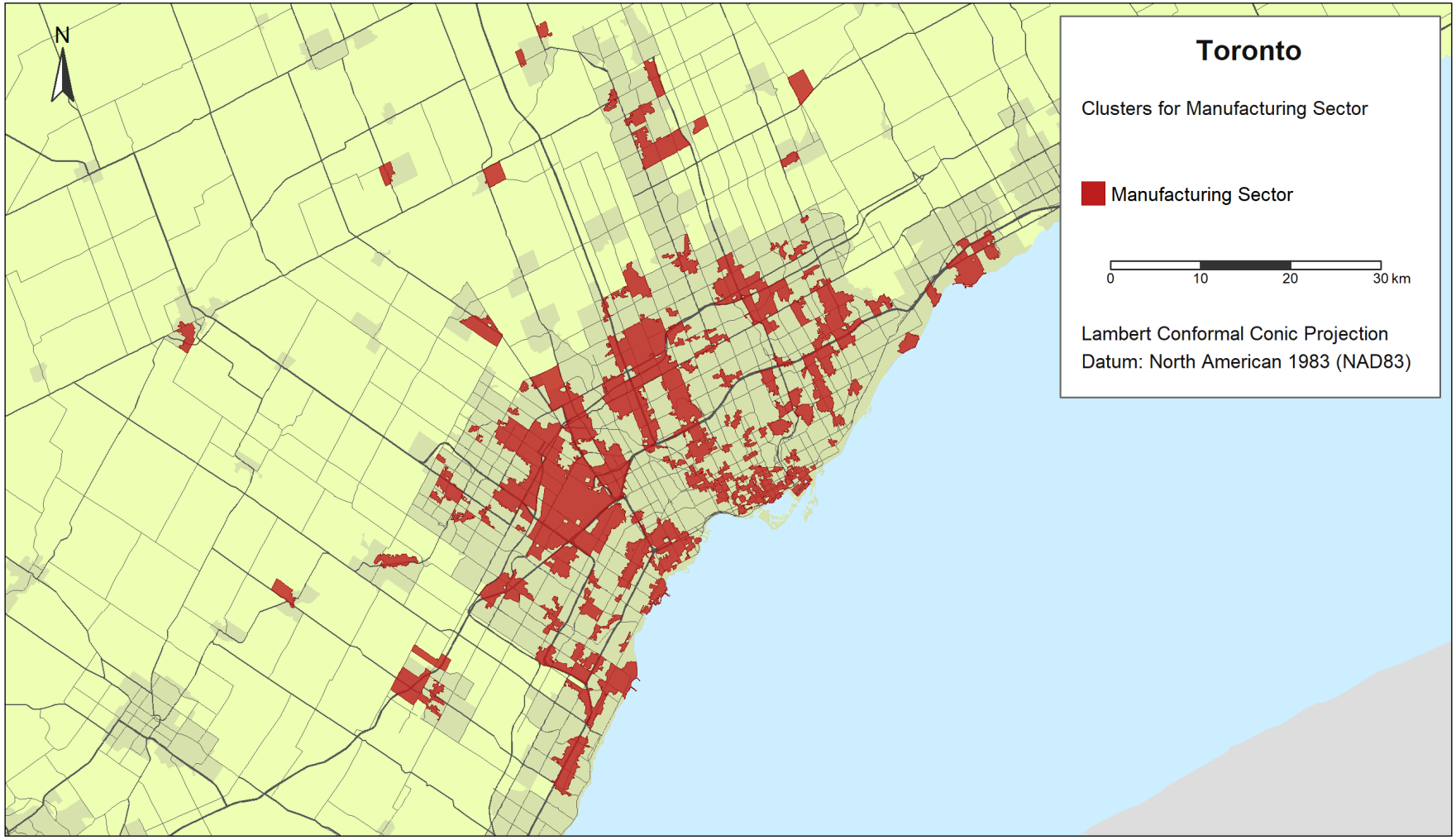
Note: 26% of DBs within the CMA, containing 89% of the employees in this industry sector
Sources: Statistics Canada, 2021 Census – Provinces/territories boundary file, 2021 Census – Population centres boundary file, 2021 Census – Road network file, 2016 Census - Coastal waters boundary file, 2016 Census - Lakes and Rivers boundary file, and authors' computations from the BR database.

Map 10
Montréal Hospitality and Tourism (cluster 22)



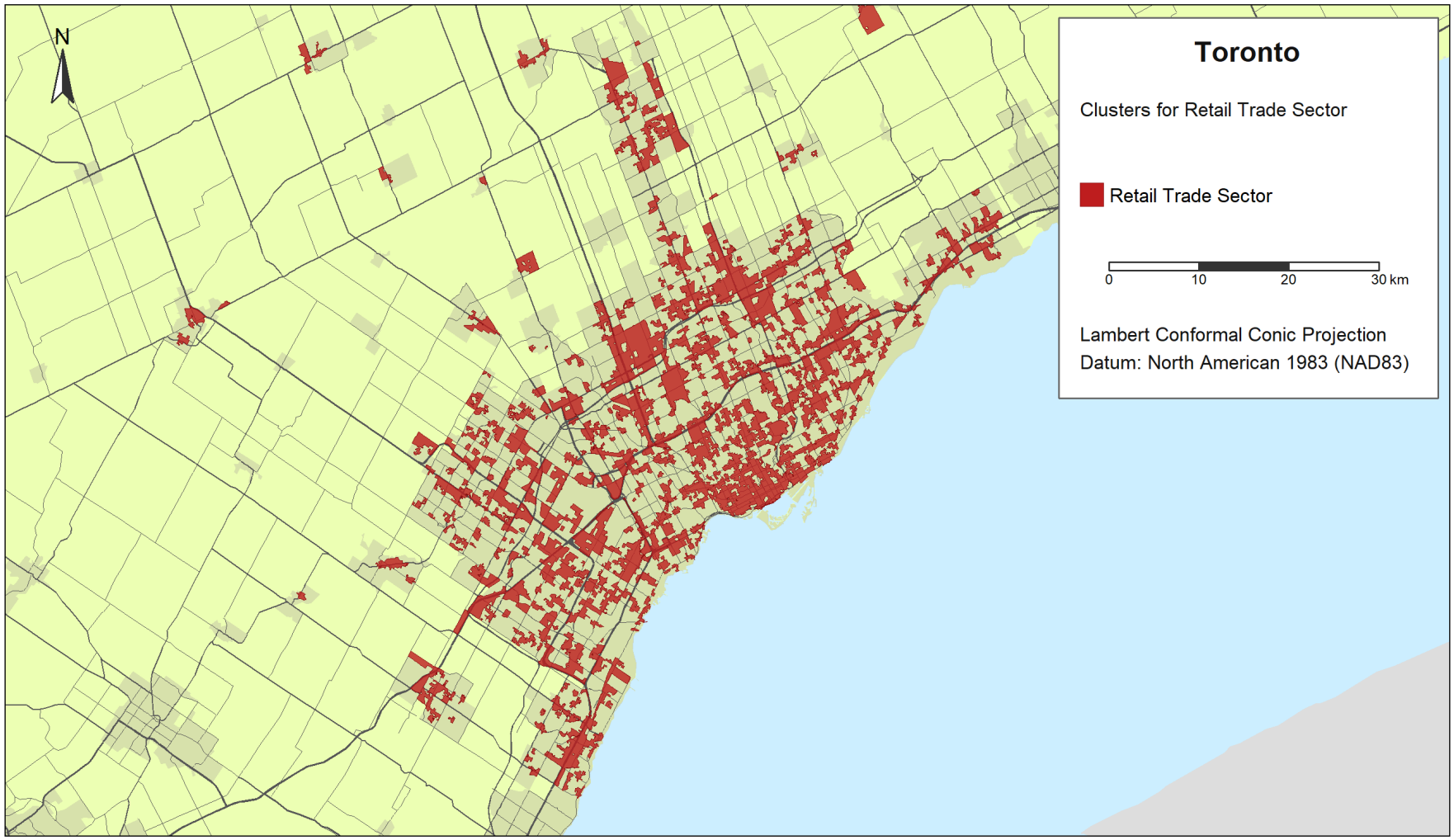
Note: 24% of DBs within the CMA, containing 60% of the employees in this industry sector
Sources: Statistics Canada, 2021 Census – Provinces/territories boundary file, 2021 Census – Population centres boundary file, 2021 Census – Road network file, 2016 Census - Coastal waters boundary file, 2016 Census - Lakes and Rivers boundary file, and authors' computations from the BR database.

Map 11
Toronto Manufacturing Sector



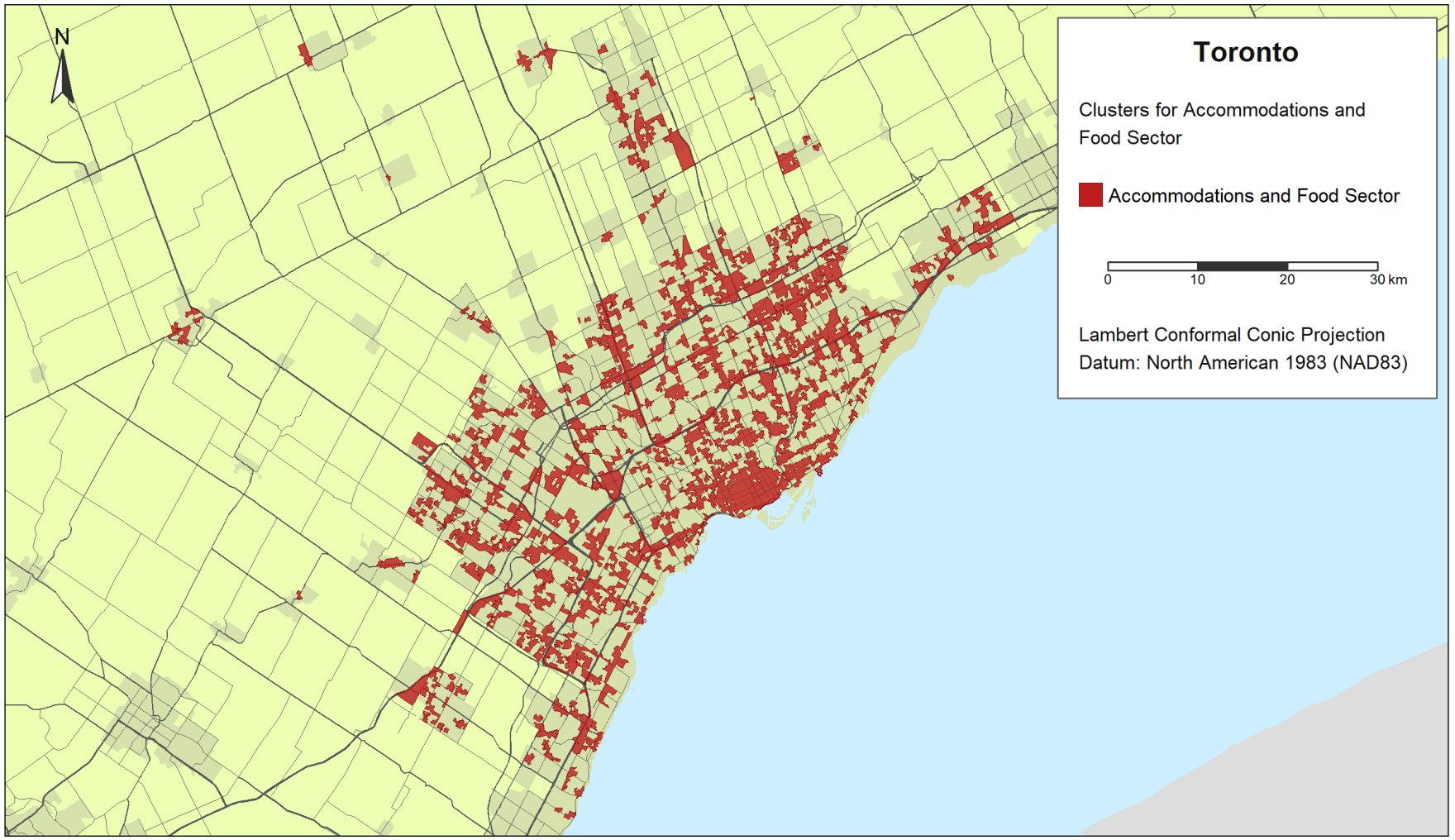
Note: 34% of DBs within the CMA, containing 94% of the employees in this industry sector
Sources: Statistics Canada, 2021 Census – Provinces/territories boundary file, 2021 Census – Population centres boundary file, 2021 Census – Road network file, 2016 Census - Coastal waters boundary file, 2016 Census - Lakes and Rivers boundary file, and authors' computations from the BR database.

Map 12
Toronto Retail Trade Sector



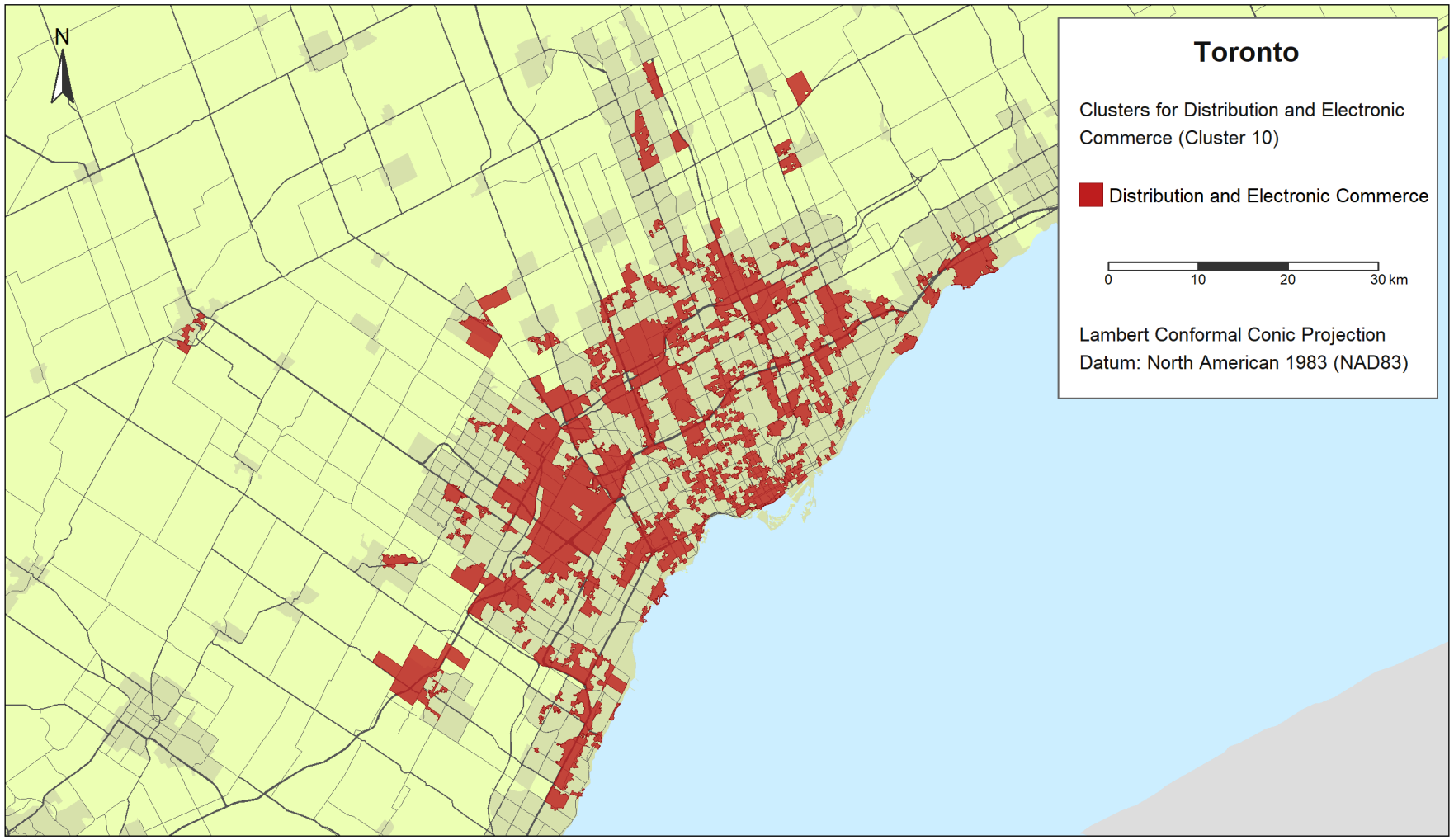
Note: 35% of DBs within the CMA, containing 97% of the employees in this industry sector
Sources: Statistics Canada, 2021 Census – Provinces/territories boundary file, 2021 Census – Population centres boundary file, 2021 Census – Road network file, 2016 Census - Coastal waters boundary file, 2016 Census - Lakes and Rivers boundary file, and authors' computations from the BR database.

Map 13
Toronto Accommodations and Food Services Sector



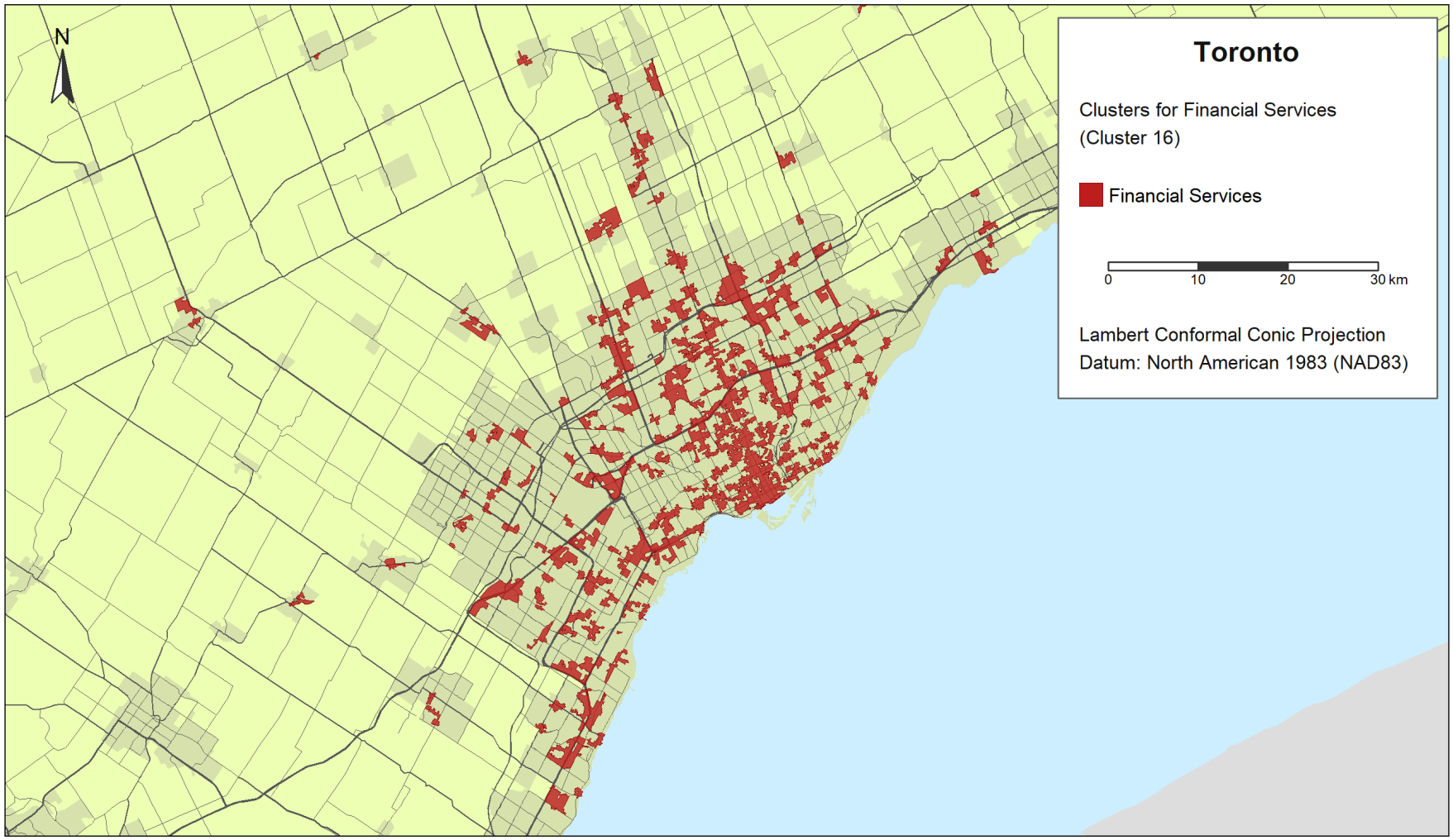
Note: 38% of DBs within the CMA, containing 89% of the employees in this industry sector
Sources: Statistics Canada, 2021 Census – Provinces/territories boundary file, 2021 Census – Population centres boundary file, 2021 Census – Road network file, 2016 Census - Coastal waters boundary file, 2016 Census - Lakes and Rivers boundary file, and authors' computations from the BR database.

Map 14
Toronto Distribution and Electronic Commerce (cluster 10)



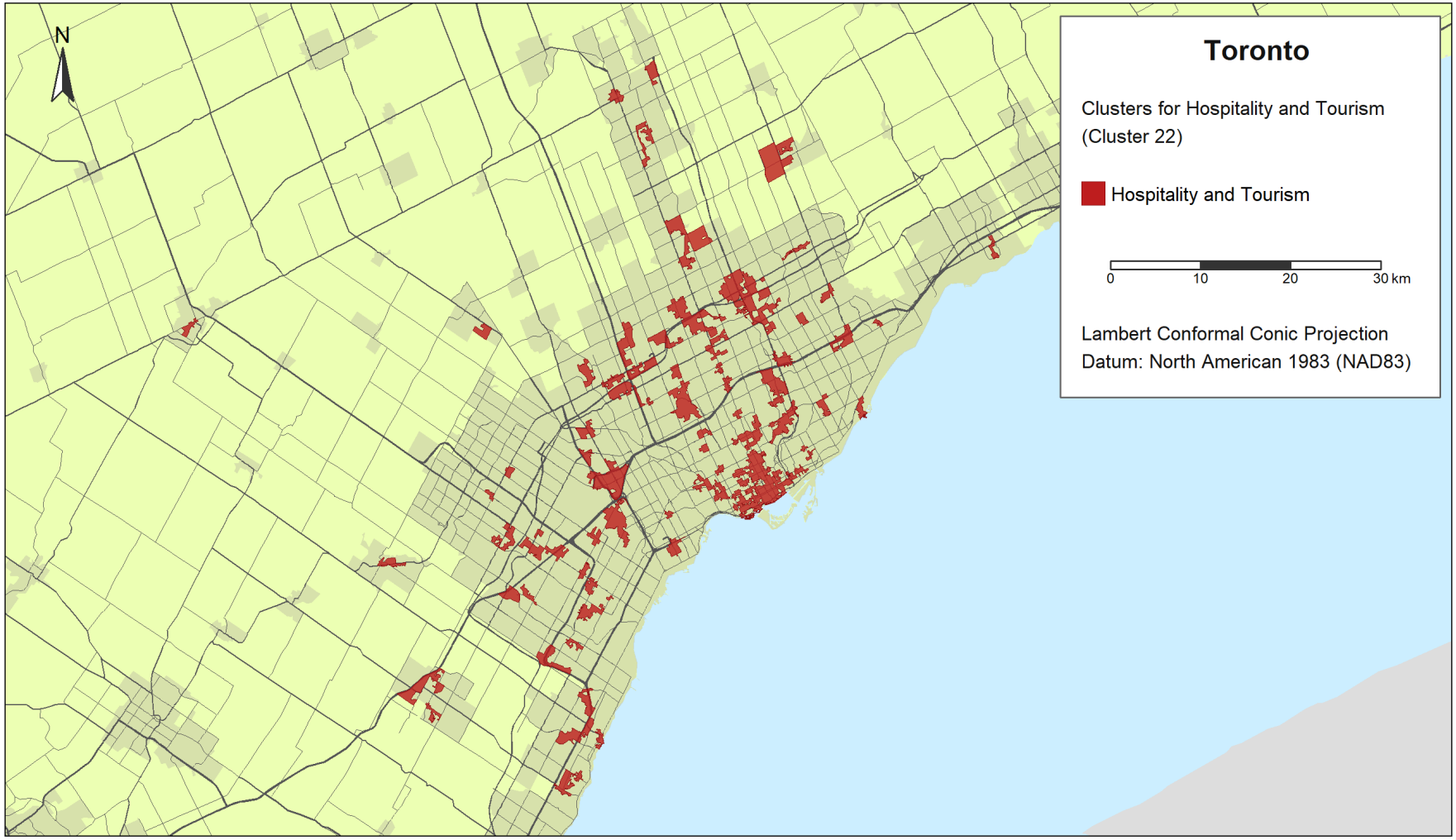
Note: 36% of DBs within the CMA, containing 98% of the employees in this industry sector
Sources: Statistics Canada, 2021 Census – Provinces/territories boundary file, 2021 Census – Population centres boundary file, 2021 Census – Road network file, 2016 Census - Coastal waters boundary file, 2016 Census - Lakes and Rivers boundary file, and authors' computations from the BR database.

Map 15
Toronto Financial Services (cluster 16)



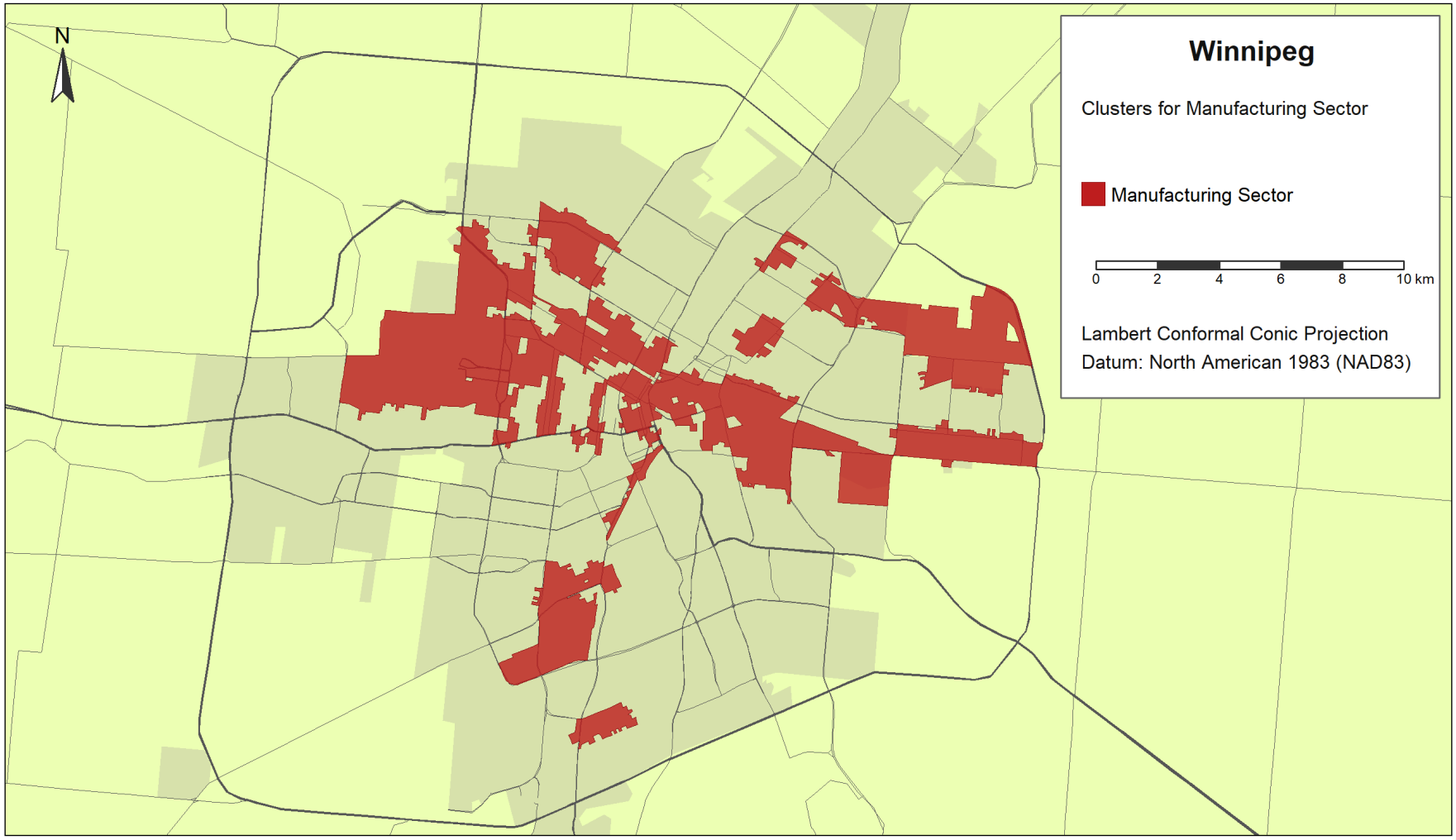
Note: 36% of DBs within the CMA, containing 96% of the employees in this industry sector
Sources: Statistics Canada, 2021 Census – Provinces/territories boundary file, 2021 Census – Population centres boundary file, 2021 Census – Road network file, 2016 Census - Coastal waters boundary file, 2016 Census - Lakes and Rivers boundary file, and authors' computations from the BR database.

Map 16
Toronto Hospitality and Tourism (cluster 22)



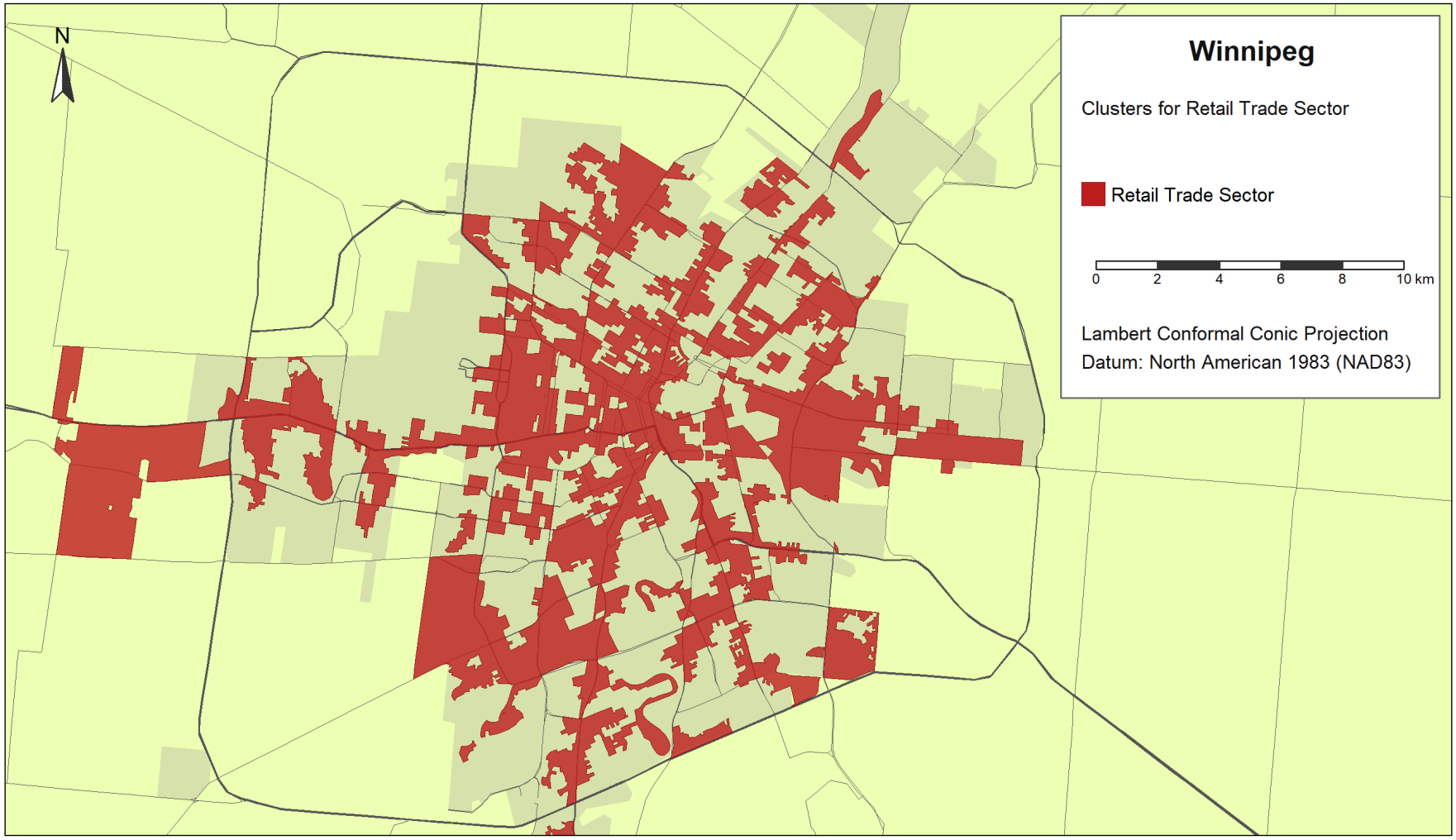
Note: 25% of DBs within the CMA, containing 71% of the employees in this industry sector
Sources: Statistics Canada, 2021 Census – Provinces/territories boundary file, 2021 Census – Population centres boundary file, 2021 Census – Road network file, 2016 Census - Coastal waters boundary file, 2016 Census - Lakes and Rivers boundary file, and authors' computations from the BR database.

Map 17
Winnipeg Manufacturing Sector



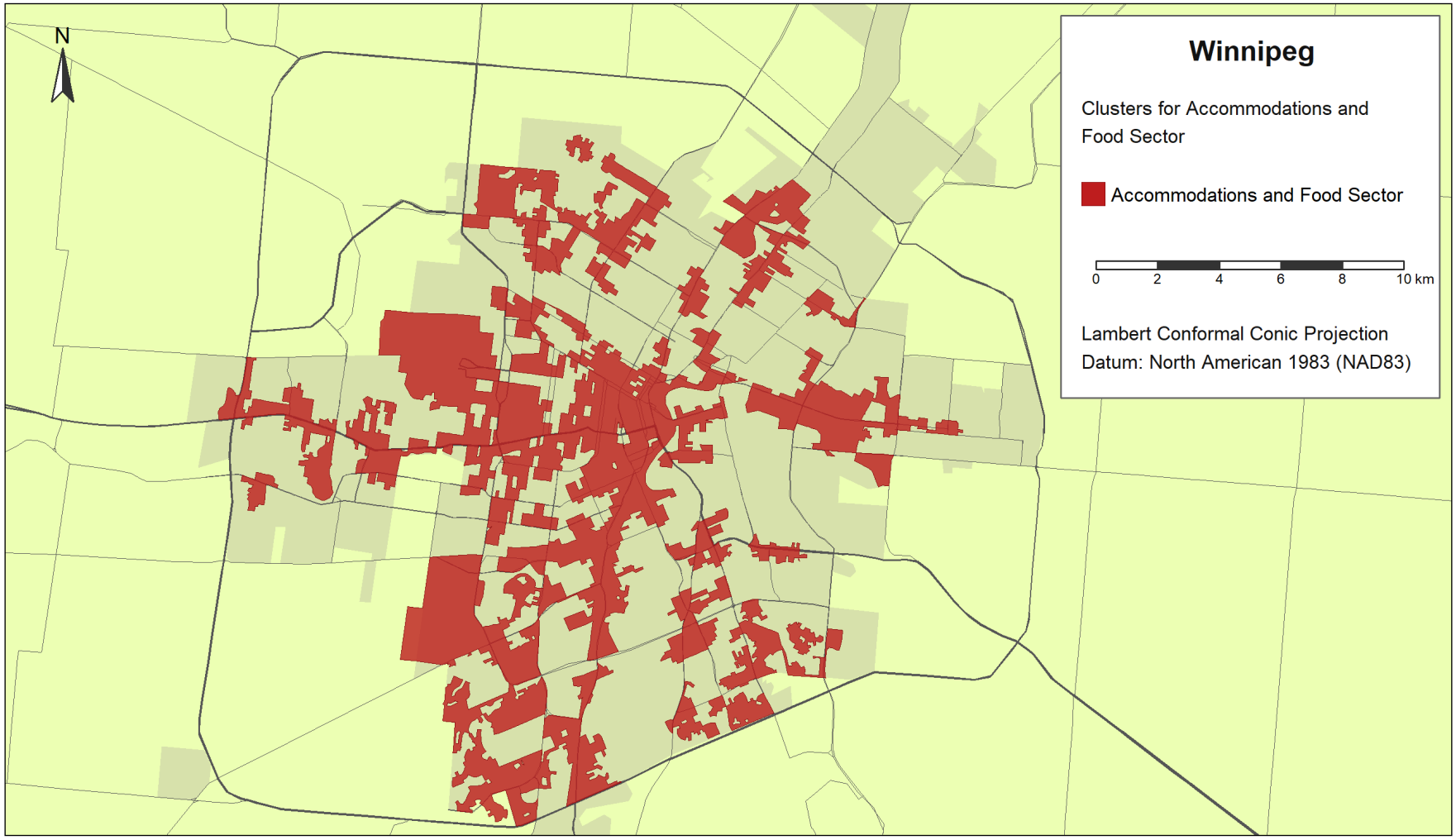
Note: 44% of DBs within the CMA, containing 91% of the employees in this industry sector
Sources: Statistics Canada, 2021 Census – Provinces/territories boundary file, 2021 Census – Population centres boundary file, 2021 Census – Road network file, 2016 Census - Lakes and Rivers boundary file, and authors' computations from the BR database.

Map 18
Winnipeg Retail Trade Sector



Note: 55% of DBs within the CMA, containing 94% of the employees in this industry sector
Sources: Statistics Canada, 2021 Census – Provinces/territories boundary file, 2021 Census – Population centres boundary file, 2021 Census – Road network file, 2016 Census - Lakes and Rivers boundary file, and authors' computations from the BR database.

Map 19
Winnipeg Accommodations and Food Services Sectorv

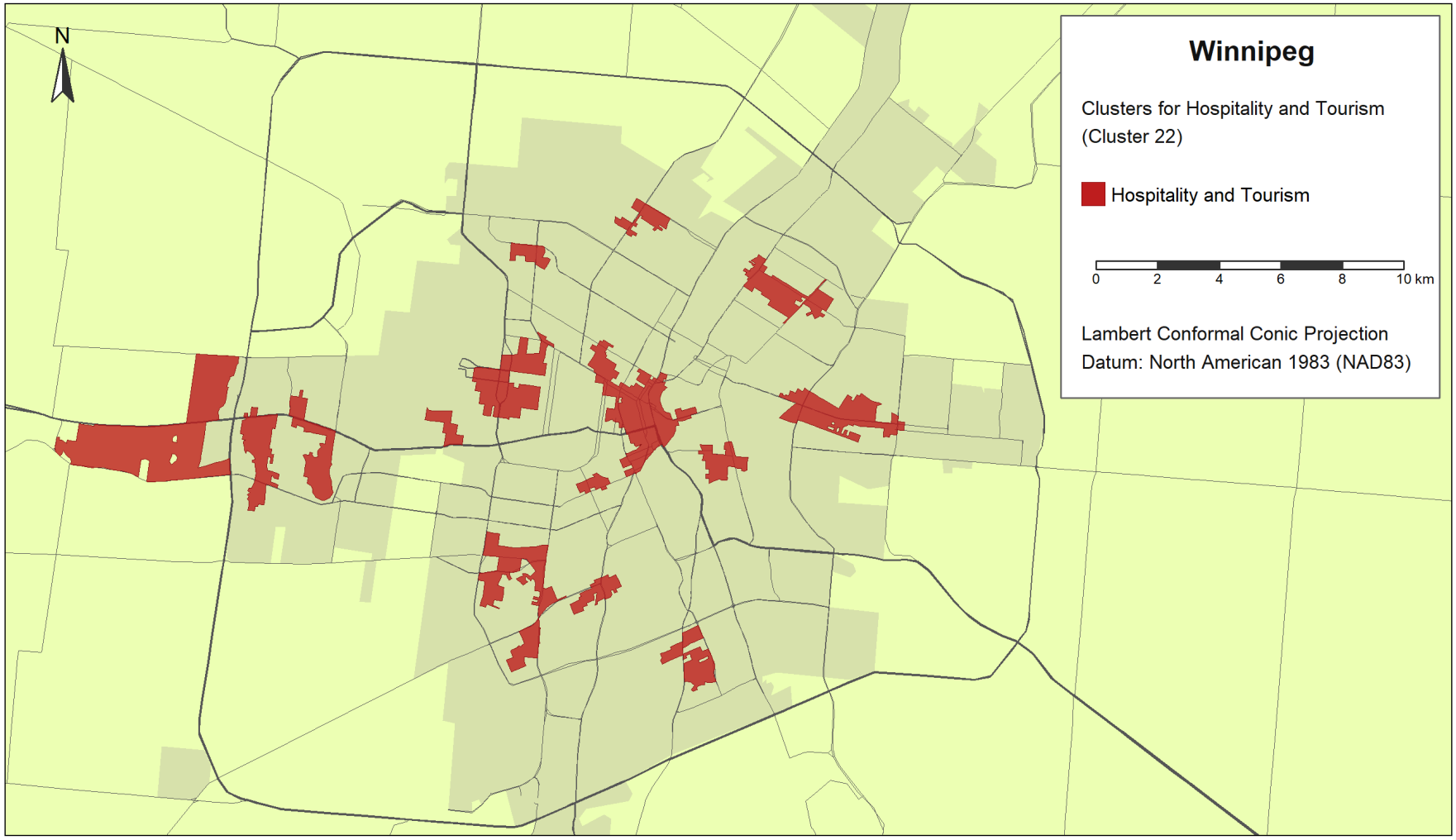


Note: 54% of DBs within the CMA, containing 91% of the employees in this industry sector
Sources: Statistics Canada, 2021 Census – Provinces/territories boundary file, 2021 Census – Population centres boundary file, 2021 Census – Road network file, 2016 Census - Lakes and Rivers boundary file, and authors' computations from the BR database.

Map 20
Winnipeg Distribution and Electronic Commerce (cluster 10)

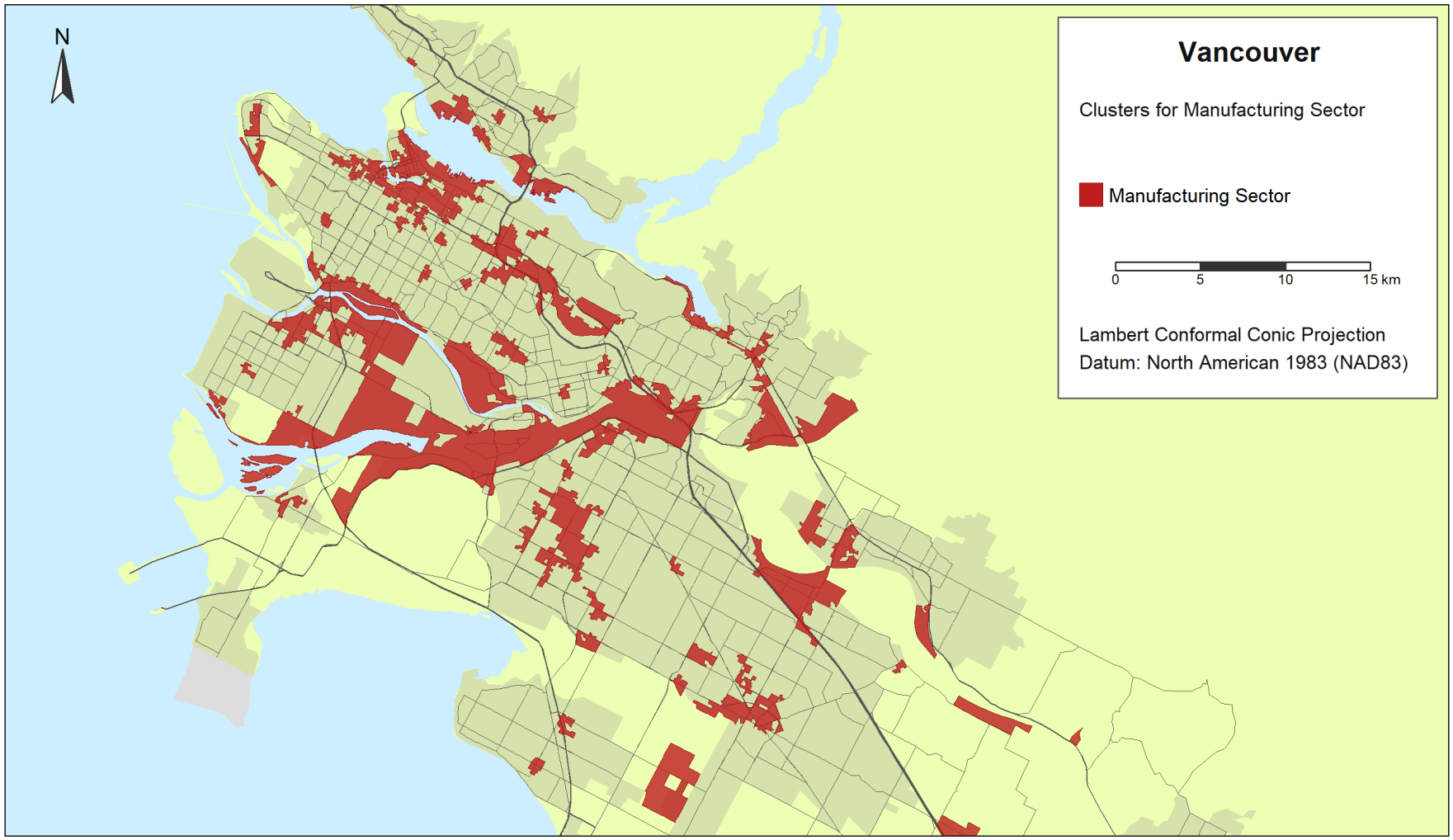
Note: 33% of DBs within the CMA, containing 93% of the employees in this industry sector
Sources: Statistics Canada, 2021 Census – Provinces/territories boundary file, 2021 Census – Population centres boundary file, 2021 Census – Road network file, 2016 Census - Lakes and Rivers boundary file, and authors' computations from the BR database.

Map 22
Winnipeg Hospitality and Tourism (cluster 22)



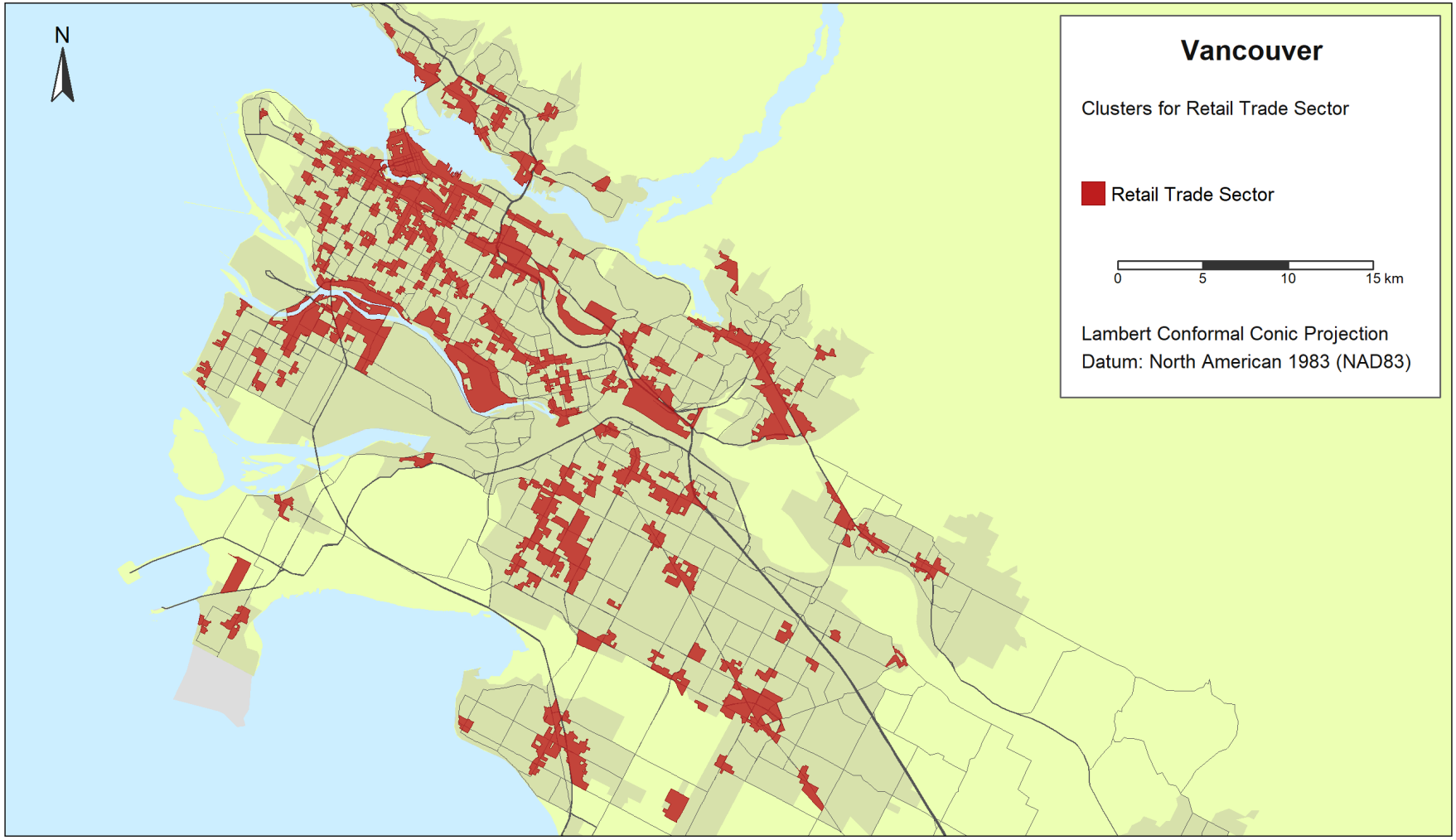
Note: 27% of DBs within the CMA, containing 60% of the employees in this industry sector
Sources: Statistics Canada, 2021 Census – Provinces/territories boundary file, 2021 Census – Population centres boundary file, 2021 Census – Road network file, 2016 Census - Lakes and Rivers boundary file, and authors' computations from the BR database.

Map 23
Vancouver Manufacturing Sector



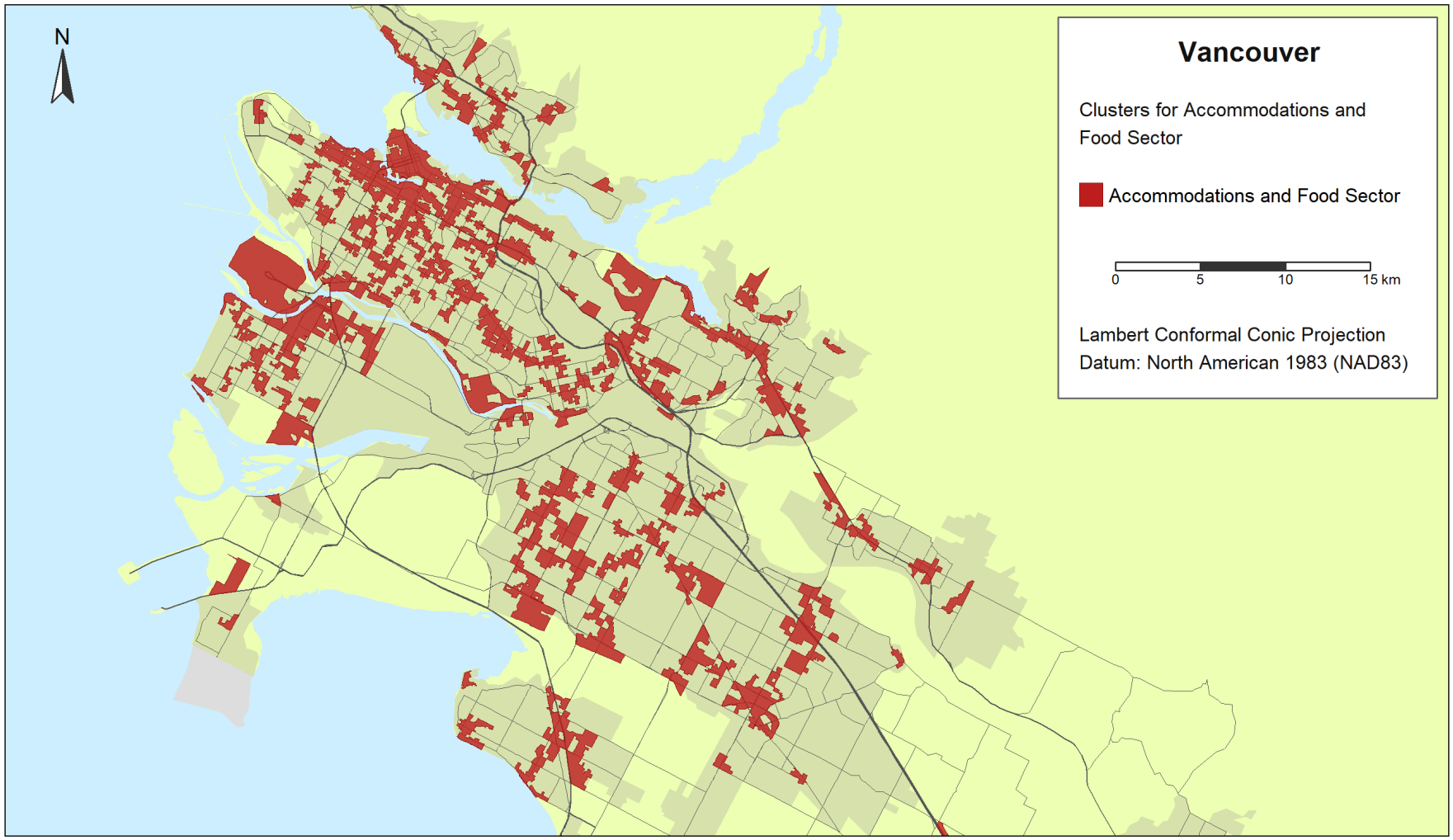
Note: 35% of DBs within the CMA, containing 90% of the employees in this industry sector
Sources: Statistics Canada, 2021 Census – Provinces/territories boundary file, 2021 Census – Population centres boundary file, 2021 Census – Road network file, 2016 Census - Coastal waters boundary file, 2016 Census - Lakes and Rivers boundary file, and authors' computations from the BR database.

Map 24
Vancouver Retail Trade Sector



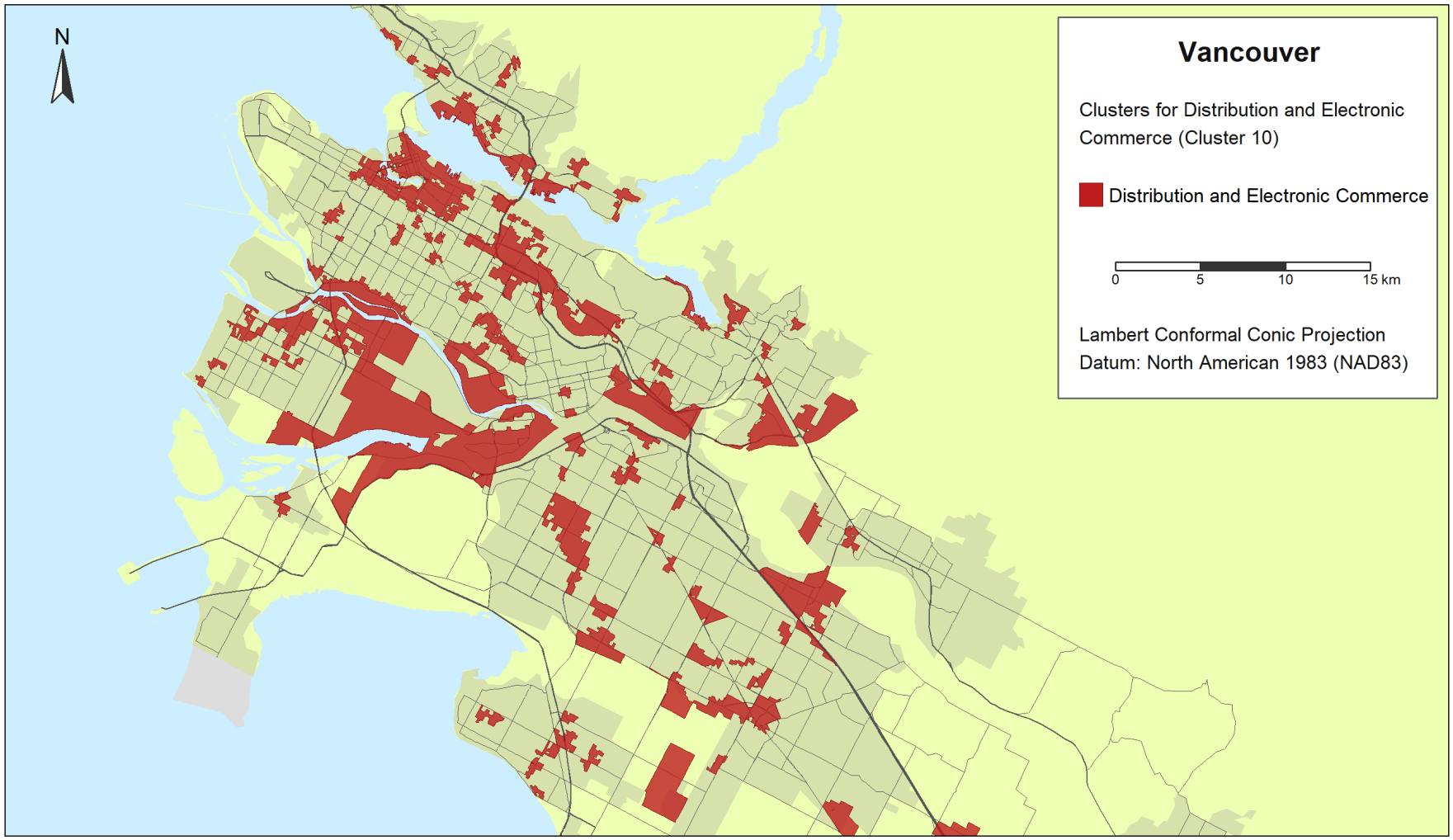
Note: 36% of DBs within the CMA, containing 81% of the employees in this industry sector
Sources: Statistics Canada, 2021 Census – Provinces/territories boundary file, 2021 Census – Population centres boundary file, 2021 Census – Road network file, 2016 Census - Coastal waters boundary file, 2016 Census - Lakes and Rivers boundary file, and authors' computations from the BR database.

Map 25
Vancouver Accommodations and Food Services Sector



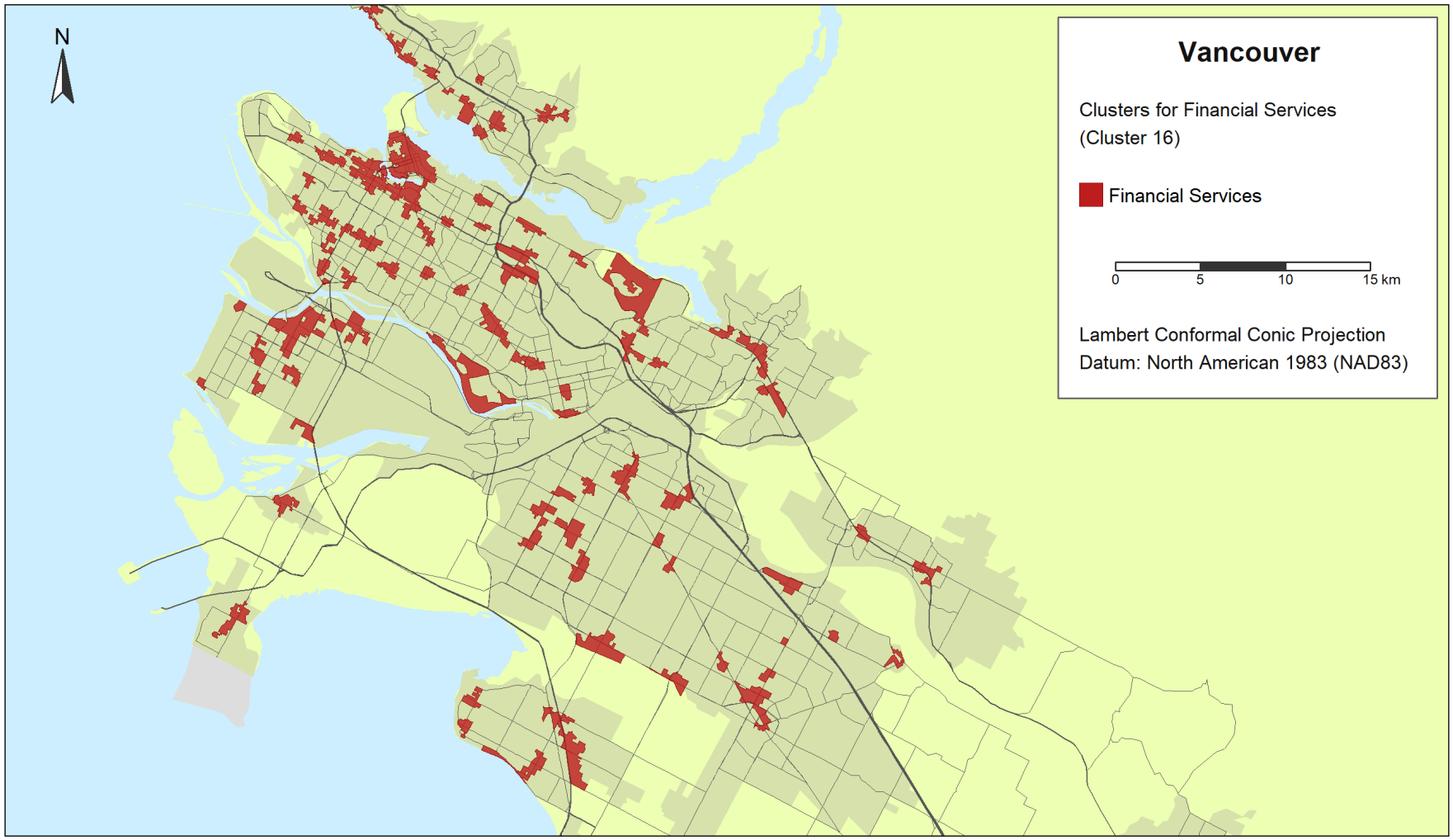
Note: 43% of DBs within the CMA, containing 91% of the employees in this industry sector
Sources: Statistics Canada, 2021 Census – Provinces/territories boundary file, 2021 Census – Population centres boundary file, 2021 Census – Road network file, 2016 Census - Coastal waters boundary file, 2016 Census - Lakes and Rivers boundary file, and authors' computations from the BR database.

Map 26
Vancouver Distribution and Electronic Commerce (cluster 10)



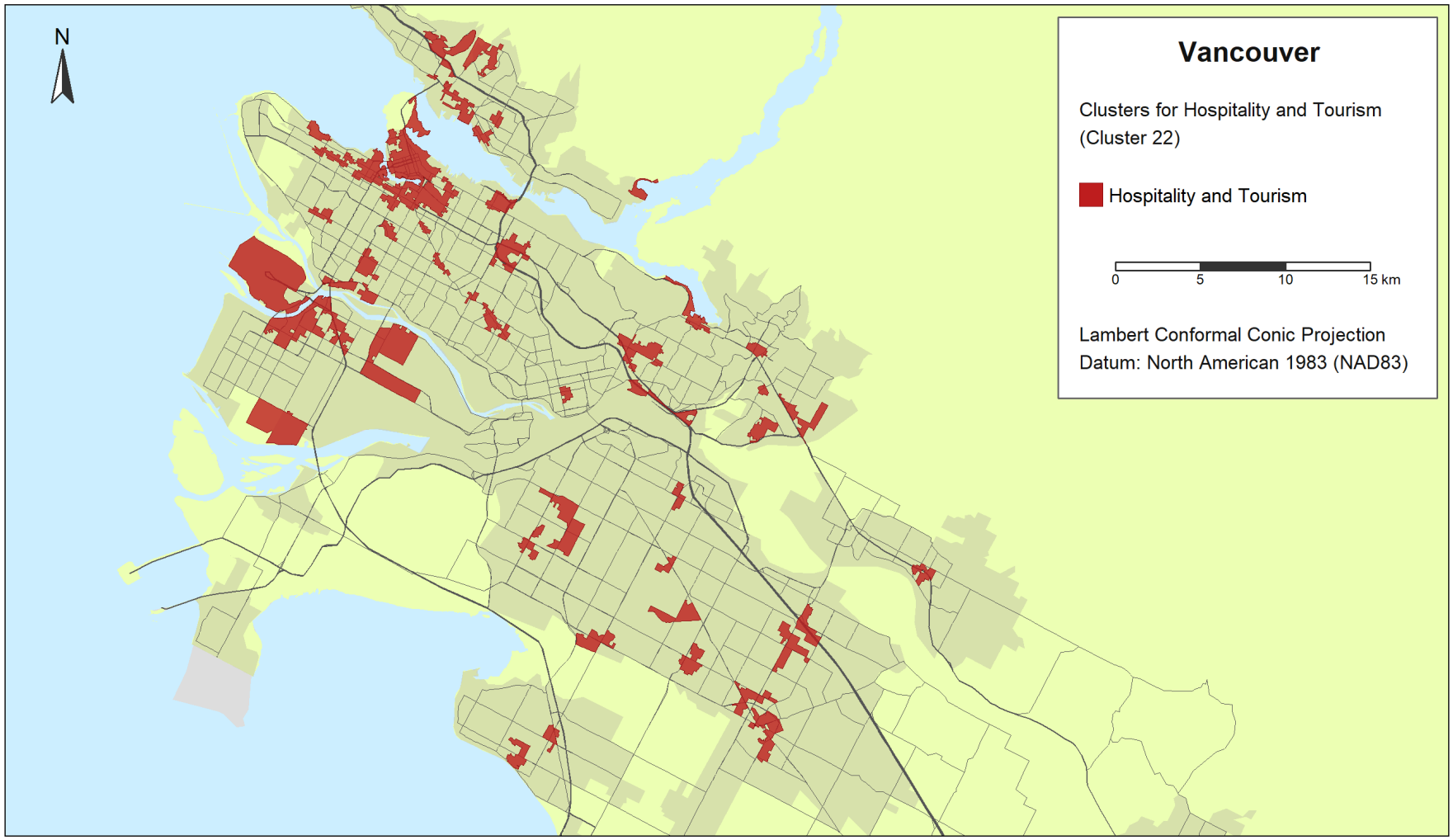
Note: 32% of DBs within the CMA, containing 94% of the employees in this industry sector
Sources: Statistics Canada, 2021 Census – Provinces/territories boundary file, 2021 Census – Population centres boundary file, 2021 Census – Road network file, 2016 Census - Coastal waters boundary file, 2016 Census - Lakes and Rivers boundary file, and authors' computations from the BR database.

Map 27
Vancouver Financial Services (cluster 16)



Note: 32% of DBs within the CMA, containing 91% of the employees in this industry sector
Sources: Statistics Canada, 2021 Census – Provinces/territories boundary file, 2021 Census – Population centres boundary file, 2021 Census – Road network file, 2016 Census - Coastal waters boundary file, 2016 Census - Lakes and Rivers boundary file, and authors' computations from the BR database.

Map 28
Vancouver Hospitality and Tourism (cluster 22)



Note: 31% of DBs within the CMA, containing 80% of the employees in this industry sector
Sources: Statistics Canada, 2021 Census – Provinces/territories boundary file, 2021 Census – Population centres boundary file, 2021 Census – Road network file, 2016 Census - Coastal waters boundary file, 2016 Census - Lakes and Rivers boundary file, and authors' computations from the BR database.