IFC-Bank of Italy Workshop on "Machine learning in central banking"

19-22 October 2021, Rome / virtual event

Cloud computing research collaboration: an application to access to cash and financial services[1]

Danielle V Handel (Stanford Institute for Economic Policy Research, Stanford University),
Anson T Y Ho (Ted Rogers School of Management, Toronto Metropolitan University),
Kim P Huynh (Bank of Canada), David T Jacho-Chavez and Carson Rea (Emory University)

[1] This presentation was prepared for the Workshop. The views expressed are those of the authors and do not necessarily reflect the views of the Bank of Italy, the BIS, the IFC or the central banks and other institutions represented at the event.

Cloud Computing Research Collaboration: An Application to Access to Cash and Financial Services*

Danielle V. Handel† Anson T. Y. Ho‡ Kim P. Huynh§ David T. Jacho-Chávez¶
Carson H. Rea‖

Abstract

We illustrate the utility of cloud computing tools for big data management and analysis serving the functions of the Bank of Canada. These tools provide the opportunity to easily leverage increasingly complex and large-scale data in an interactive coding environment without worrying about backend infrastructure. As an empirical use case to demonstrate these advantages, we use a cloud computing platform to expedite a computationally intensive spatial analysis mapping access to financial services in Canada.

Keywords: High-Performance Computing; Big data; Spark; Jupyter.

JEL codes: A11; A22; A23; C87; C88.

*We are grateful to Giuseppe Bruno for his helpful comments and suggestions. We also thank Brian Nadon, Phil Riopelle and the Analytical Environment Business Systems team at the Bank of Canada for their excellent assistance in the Digital Analytical Zone. The views expressed in this article are those of the authors. No responsibility for them should be attributed to the Bank of Canada. All remaining errors are the responsibility of the authors.

†Stanford Institute for Economic Policy Research, Stanford University, John A. and Cynthia Fry Gunn Building, 366 Galvez St, Stanford, CA 94305, USA. E-mail: dvhandel@stanford.edu

‡Ted Rogers School of Management, Toronto Metropolitan University, 55 Dundas Street West, Toronto ON, M5G 2C3, Canada. E-mail: atyho@ryerson.ca

§Corresponding Author: Bank of Canada, 234 Wellington Ave., Ottawa ON, K1A 0G9, Canada. E-mail: khuynh@bankofcanada.ca.

¶Department of Economics, Emory University, Rich Building 306, 1602 Fishburne Dr., Atlanta,GA 30322-2240, USA. E-mail: djachocha@emory.edu

‖Department of Economics, Emory University, Rich Building 306, 1602 Fishburne Dr., Atlanta, GA 30322-2240, USA. E-mail: chrea@emory.edu

1 Introduction

Demand for computational resources have been rapidly expanding in recent years, driven by the increasing interest in big data, new analytical methods in data science, and a slowdown in technological progress. In response, cloud computing has become a popular solution for institutions to meet their computational needs. At the Bank of Canada, the *Digital Analytical Zone* (DAZ) is a cloud computing platform implemented through Microsoft® Azure, their contracted vendor. It allows for an *on-demand* computing service that is agile and highly scalable to meet the resource requirement of different projects, aligning resources with needs and making high performance computing more accessible.

This case study illustrates the use of the Bank of Canada's DAZ to access big data tools for research collaboration with external academic researchers. Our data analysis is conducted on Azure Databricks, an Apache Spark-based data analytics service. As shown in Handel et al. (2021), cloud computing is convenient as it removes infrastructure constraints. However, depending on the application complexity, setting up virtual machines on the cloud may still require considerable initial cost and expertise in cloud computing. Managed cloud service platform simplifies the use of cloud computing by providing a pre-configured computational service. In our case, Azure Databricks provides a fully managed Spark cluster that enable researchers to easily harness the power of data parallelism for large scale data processing.

The Bank of Canada's DAZ responds to the challenges and options highlighted in Bruno et al. (2020). It shows a promising path to efficiently employ big data analysis. This cloud-based platform provides the opportunity to easily leverage the increasingly complex "financial big data sets" and work with innovative data to yield new insights important to the functions of central banks. For example, Canadian consumer credit data is used to analyze the effect of COVID-19 on consumer finance (Ho, 2020) and the interdependence of financial institutions in the consumer credit markets (Ho et al., 2021). For the rest of the paper, Section 2 describes the cloud computing service at the Bank of Canada. Section 3 provides a use case of cloud computing, and Section 4 concludes.

2 Cloud Computing at the Bank of Canada

The Bank of Canada currently offers computational resources through several different channels. Researchers can access an on-premise high-performance-computing (HPC) cluster named *Edith2*.[1] In addition, the Bank of Canada also provides a cloud computing environment called the *Digital Analytical Zone (DAZ)* to support scientific research (Elsey et al., 2021). In general, the DAZ is designed to be completely separate from the Bank of Canada's network. It provides an environment for users to experiment with different ideas. The DAZ is supported by Microsoft Azure®, the Bank of Canada's service vendor, which provides various types of cloud computing services. *Research Services* is a managed cloud computing platform, where users can launch their virtual machines with pre-configured specifications. Based on business needs, the DAZ also provide a

[1] For more information, see Collignon (2019).

more flexible *Research Lab* platform, closer to full-fledged Microsoft Azure®, where users can configure their virtual machines.

The DAZ offers several advantages over an on-premise HPC cluster. First and foremost, it provides users with *on-demand* computing, with which users do not have to wait in queue due to Edith2's capacity limit. Timely access to a computational resource increases users' productivity, particularly when there is expanding demand from big data and more complex data science methods. It also allows users to access a computational resource on the internet without the need of pre-configured physical device. Second, the DAZ is more flexible in providing up-to-date services than an on-premise HPC cluster. While most of the cutting-edge data science methods come from community-contributed libraries, installing them on an on-premise HPC cluster often requires extensive testing and system configuration. On the other hand, the DAZ is maintained by the service vendor, which benefit from the economies of scale and affords users to have a higher level of administrative rights. This flexibility encourages users to explore new techniques and promotes innovation. Third, the DAZ can easily extend the Bank of Canada's computational resource to external partners for project collaboration. DAZ administrators can simply create accounts for external partners on the cloud platform for instant collaboration, instead of granting external partners access to the on-premise HPC cluster that may involve costly equipment and lengthy security clearance.

In this paper, we demonstrate how the DAZ can facilitate collaborations between the Bank of Canada researchers and external partners. While the interface and the functions available on the DAZ are identical to a typical Microsoft Azure® portal, some functionality requires prior approval from the DAZ administrator. To initialize a project, a DAZ administrator sets up a resource group on Azure according to the user's business case. If external collaborators are involved, additional accounts are created and assigned to that specific resource group.[2] All users within a resource group shares the same pool of computational resource. For big data analysis and machine learning, we further focus on Azure Databricks, among other services available.

2.1 Azure Databricks

Azure Databricks is an Apache Spark-based data analytics service. It supports multiple programming languages, including Python, Scala, R and SQL. It also supports popular machine learning libraries such as Apache Spark MLlib, Tensorflow, Pytorch, allowing for the use of both advanced statistical and machine learning techniques. Its markdown-compatible notebook environment also provides the opportunity for clear documentation, efficient debugging, and promoting reproducible research.

After entering the Azure portal, utilizing these tools involves first creating and naming a Databricks resource. Researchers can then launch a Databricks workspace, which is an interactive interface for managing all of the Databricks tools. Figure 1 shows the layout of Azure Databricks. Within the workspace, there are various options for launching a project-specific cluster, which is a collection of servers that will provide the

[2]The DAZ's Azure account is independent of other Microsoft Azure accounts that a user may have.

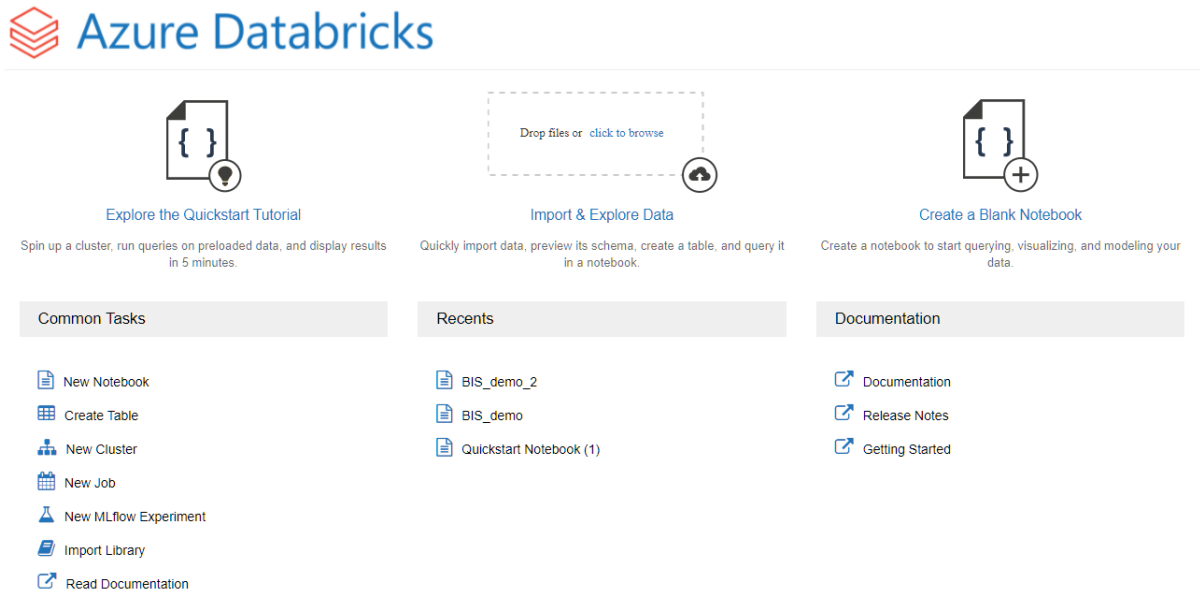


Figure 1: Screenshot of Microsoft Azure Databricks

computing power needed to complete big data tasks. For the demonstration in this paper, we provision a standard cluster with 14 GB of memory and 4 cores. Memory size and parallel computing specifications are highly customizable, with options for additional memory and cores and different runtimes. Also available is the serverless option that allows for auto-scaling, meaning that the resources employed will be used efficiently and adjusted upon need. As opposed to shared on-premise HPC cluster, these on-demand project-specific clusters can be deployed without going through a job scheduler. After initiating a cluster, any project data can be uploaded into the Databricks File System (DBFS), converted into a table using the user interface, and subsequently used in any projects in the workspace. Researchers may refer to any data stored in DBFS using absolute file paths as if using data stored on a local machine.

In order to execute analyses using big data tools and interact with Apache Spark, researchers may use the interactive Databricks notebooks. After either uploading an existing notebook or creating a new Databricks notebook, attaching the notebook to a running cluster will allow for interactive management of Spark resources within the Jupyter-style interface. For the demonstration, we interact with Spark within a Python script using the PySpark interface. The Databricks notebooks provide detailed metrics regarding Spark performance and runtime, and researchers can interact with tables and other objects to quickly produce plots and summary statistics or save output to their local machine. Researchers may choose to link Databricks notebooks to Github repositories using built-in Git integration, which allows for version control working with notebook revision histories and enhanced capability for collaboration. The built-in Git integration also includes the capability to create branches and pull requests in the relevant repository from within the Databricks UI.

3 Application – Canadian Access to Financial Services

In our empirical application, we measure Canadians' access to financial services by computing the distances between their residential locations and the nearest branch of a financial institution (FI) in 2017. Physical proximity as a measurement for access to financial services is supported by Mintel's (2018) survey findings that consumers rely on access to physical branches for the purchase of complex financial products or first-time banking interactions. Linking a spatial network of FIs with methods-of-payment survey further yields a comprehensive analysis on the role of banking in influencing consumer payment choices (Henry et al., 2018) and their cash withdrawal (Chen et al., 2021a).

To measure the physical proximity to local branches, we compute the straight-line distances from the population centroid of each postal code to all FIs, and then identify the distance to the closest branch. The same methodology is used in Tischer et al. (2020) and Chen et al. (2021b). The postal code data set comes from Statistics Canada's Postal Code Conversion File (PCCF) and the addresses of financial institutions are reported in Payment Canada's Financial Institutions File (FIF). We further appended the postal codes data set with additional demographic information at census-dissemination-area level from the 2016 Canadian Census.

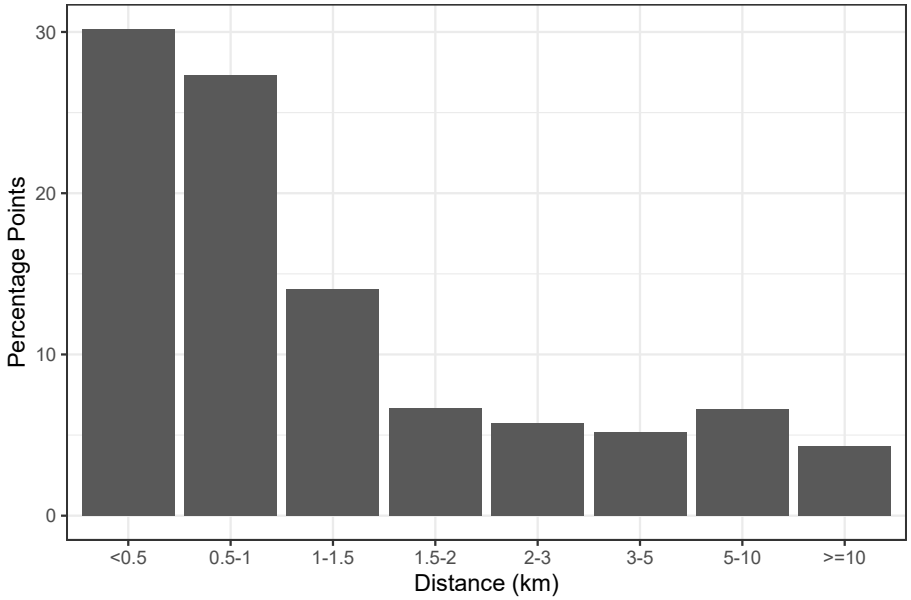
This application demonstrates the value of a cloud computing platform, by expediting a process which typically requires high local computing power and extensive time. In 2017, FIs operated a total of 11,029 local branches across Canada, serving Canadians resided in 765,723 different postal codes. To address the volume of data and complex spatial calculations, we employ Apache Spark's dimension reduction algorithms in Azure Databricks.

Our estimates on the access to financial services are illustrated in Figure 2. Overall, 57.4% of credit active Canadian residents have access a FI branch within 1 km of their residential location. The most common distance is less than 0.5 km, which contains about 30.1% of all residents. This suggests that most residents have convenient access to FI branches for financial services. While the fraction of people farther away from FI branches drops quickly, the distribution also exhibits a long tail with 4.3% residents having the nearest branch being more than 10 km away.

Canadian residents' proximity to financial services is related to FIs strategically locating their branches (Allen et al., 2008; Chen and Strathearn, 2020). As shown in Figure 3, vast majority of residents in the highest density quintile has access to a branch with 1 km. Distance to FI branches increases gradually for people living in areas with lower population density. Notably, majority of people residing in the lowest density quintile have to travel 3 km or more to visit a FI for financial service.

While we observed distinct patterns in the access to financial service, we did not find lower income groups are disadvantage in access to FI branches. Indeed, we observed the opposite. In general, residents in the lowest income quintile neighborhoods have closest access to local branches, with almost 50% of the residents having access in less than 0.5 km. Proximity to FI branches decreases steadily in higher income neighborhoods, and the distribution also becomes more skewed. Nonetheless, in the second to the fifth income

Figure 2: Overall Access to Financial Service



quintile neighborhoods, about 70% of residents have the nearest branches within 1.5 km.

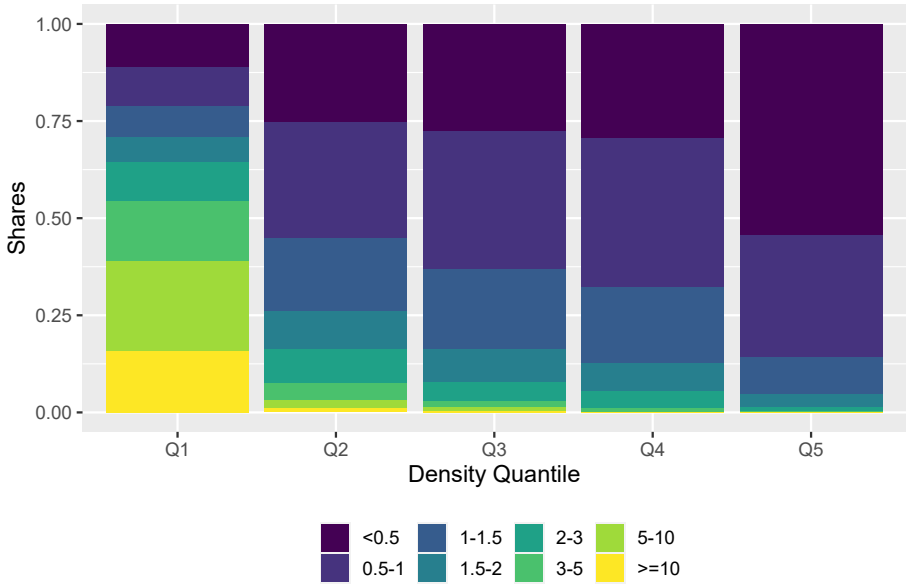
In terms of spatial pattern, we categorize postal codes into regions using the Statistical Area Classification (SAC) provided by Statistics Canada. Specifically, the land mass of Canada is divided into census metropolitan areas (CMAs), census agglomerations (CAs), and census metropolitan influenced zones (MIZs) based on population size and commuting patterns. A summary of SAC and their shares of population is reported in Table 1. Graphically, the SAC for Ottawa-Montreal-Quebec City region is illustrated in Figure 5.

Table 1: Statistical Area Classification

Classification	Description	Pop. Share
CMA	Total population $\geq$100,000; core population $\geq$50,000	0.714
CA	Total population $<$100,000; core population $\geq$10,000	0.123
Strong MIZ	$\geq$30% of employed residents commute to work in CMA or CA	0.055
Moderate MIZ	5% - 30% of employed residents commute to work in CMA or CA	0.066
Weak MIZ	0% - 5% of employed residents commute to work in CMA or CA	0.036
No MIZ	No employed residents commute to work in CMA or CA	0.005

Access to financial service exhibits a unique spatial pattern. As shown in Figure 6, residents in CMAs have the shortest distance to FI branches, since these areas have higher levels of urban development and business activities. Distance to branches increases with areas farther away from CMAs, potentially due to smaller population size. It also shows a more skewed distribution, in which some residents are very far away from any branches. For instance, about 50% of the strong MIZ residents have to travel at least 5 km for visiting a FI.

Figure 3: Access to Financial Service by Population Density



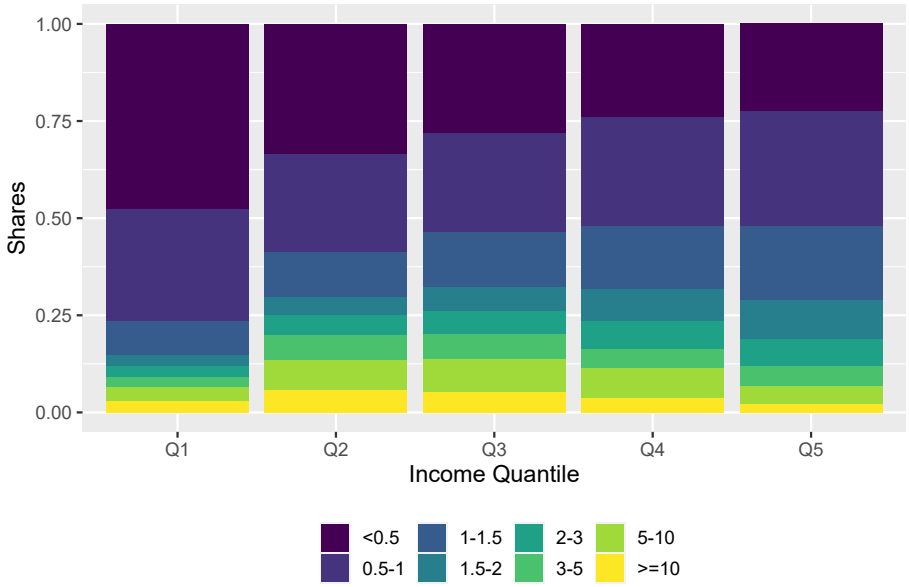
Most interestingly, distance to bank branches *decreases* from strong MIZ to weak MIZ. The fraction of residents with access in less than 1 km increases significantly from strong MIZ to weak MIZ, while that with access between 3-10 km decreases. It implies that people living outside of sizable cities have closer access to banks as the influence of metropolitan area fades. The economic explanation is that people can also access financial services via their daily trips to work. With smaller fractions of residents in moderate (5% to 30%) and weak MIZ (less than 5%) commuting to CMA or CA for work, there are stronger local banking needs that warrants the operation of a local branch. Such clustering is even more obvious in no MIZ areas, where the distribution of banking access shows a bimodal distribution with about 24% of residents having access in less than 0.5 km and about 50% of them having to travel for more than 10 km.

4 Conclusion and Considerations

We illustrate an example of how cloud computing is used at the Bank of Canada for research collaboration with external partners. Our use case shows that cloud computing is scalable and capable to handle computationally intensive data analysis. Most importantly, the cloud computing platform at the Bank of Canada allows for easy resource sharing among collaborators in different institutions, different regions, and different time zones. It abstracts out institutional-specific infrastructure configurations, providing a common platform for users to collaborate on their data analysis.

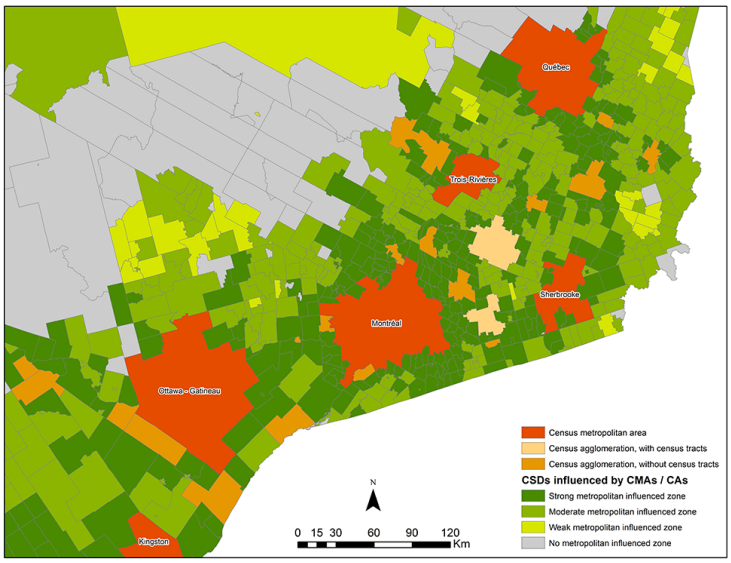
While cloud computing brings resource flexibility to institutions, implementing a cloud computing platform also involve various considerations. Migrating to a cloud platform involves training and additional support for users to adapt to a new infrastructure and a new workflow. The extra cost of time and support

Figure 4: Access to Financial Service by Income Quintile



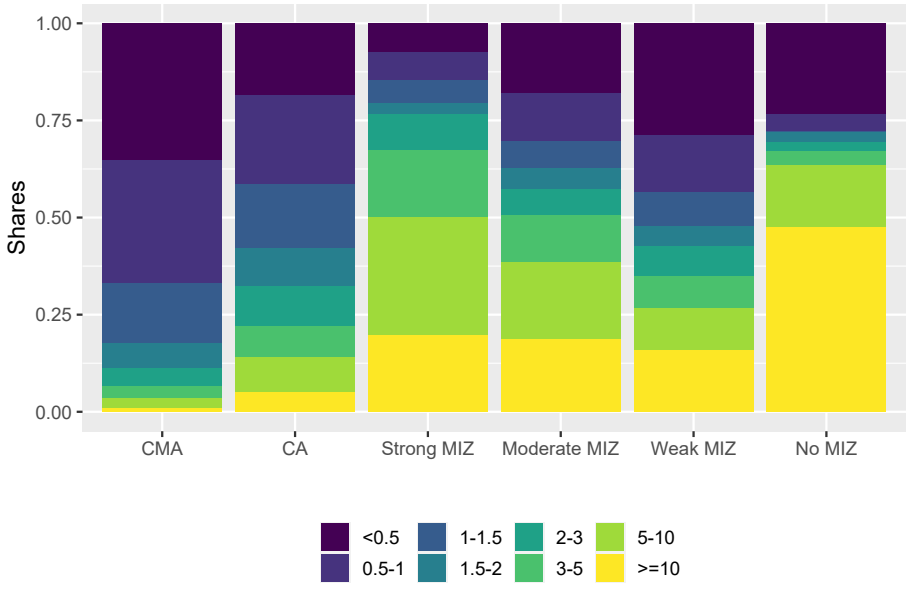
personnel should be included when comparing different solutions for fulfilling users' computational needs, at least in the short run. Furthermore, a usage policy should be set up for the implementation of cloud computing. This may entail the amount of resource and computation time budgeted for a project, as well as the type of data that can be stored on the cloud.

Figure 5: Example of Statistical Area Classification



Source: Statistics Canada, 2016 Census of Population.

Figure 6: Access to Financial Service by Area Type



References

Allen, Jason, Robert Clark, and Jean-François Houde, "Market Structure and the Diffusion of E-Commerce: Evidence from the Retail Banking Industry," Technical Report, Bank of Canada Staff Working Paper 2008-32 2008.

Bruno, Giuseppe, Hiren Jani, Rafael Schmidt, Bruno Tissot, Bank für Internationalen Zahlungsausgleich, and Irving Fisher Committee on Central Bank Statistics, *Computing platforms for big data analytics and artificial intelligence* 2020. OCLC: 1187922905.

Chen, Heng and Matthew Strathearn, "A Spatial Model of Bank Branches in Canada," Staff Working Paper 2020-4, Bank of Canada February 2020.

_ , _ , and Marcel Voia, "Consumer Cash Withdrawal Behaviour: Branch Networks and Online Financial Innovation," Technical Report, Bank of Canada Staff Working Paper 2021-28 2021.

_ , Walter Engert, Kim P. Huynh, and Daneal O'Habib, "An Exploration of First Nations Reserves and Access to Cash," Staff Discussion Paper 2021-8, Bank of Canada 2021.

Collignon, Barbara, "BOC's Analytic Environment: A Leap Into The Future," in "Bank of Italy and BIS Workshop on 'Computing Platforms for Big Data and Machine Learning"' 2019.

Elsey, Rob, Richard Harmon, Ulrika Pilestl, Ben Sorensen, and Dirk Robijns, "Central bankers at the frontier: the state of the art in advanced analytics and AI," in "BIS Innovation Summit 2021" 2021.

Handel, D.V., A.T.Y. Ho, K.P. Huynh, D.T. Jacho-Chavez, and C. Rea, "Econometrics Pedagogy and Cloud Computing: Training the Next Generation of Economists and Data Scientists," *Journal of Econometric Methods*, 2021, *10* (1), 89–102.

Henry, Christopher, Kim Huynh, and Angelika Welte, "2017 Methods-of-Payment Survey Report," *Bank of Canada Staff Discussion Papers*, 2018, (18-17).

Ho, Anson T. Y., "Interconnectedness through the Lens of Consumer Credit Markets," in Á de Paula, E Tamer, and M C Voia, eds., *The Econometrics of Networks (Advances in Econometrics, Vol. 42)*, Emerald Publishing Limited, oct 2020, pp. 315–333.

_ , Lealand Morin, Harry J. Paarsch, and Kim P. Huynh, "Consumer Credit Usage in Canada during the Coronavirus Pandemic," *Canadian Journal of Economics*, 2021, *Special issue: The COVID-19 Pandemic* (54). forthcoming.

Mintel, "The Branch Banking Experience - Canada - February 2018," Technical Report, Mintel February 2018.

Tischer, Daniel, Isobel Oxley, Jamie Evans, and Richard Scott, "Where to Withdraw: National Mapping of Access to Cash," December 2020.

BANK OF CANADA
BANQUE DU CANADA

EMORY
UNIVERSITY

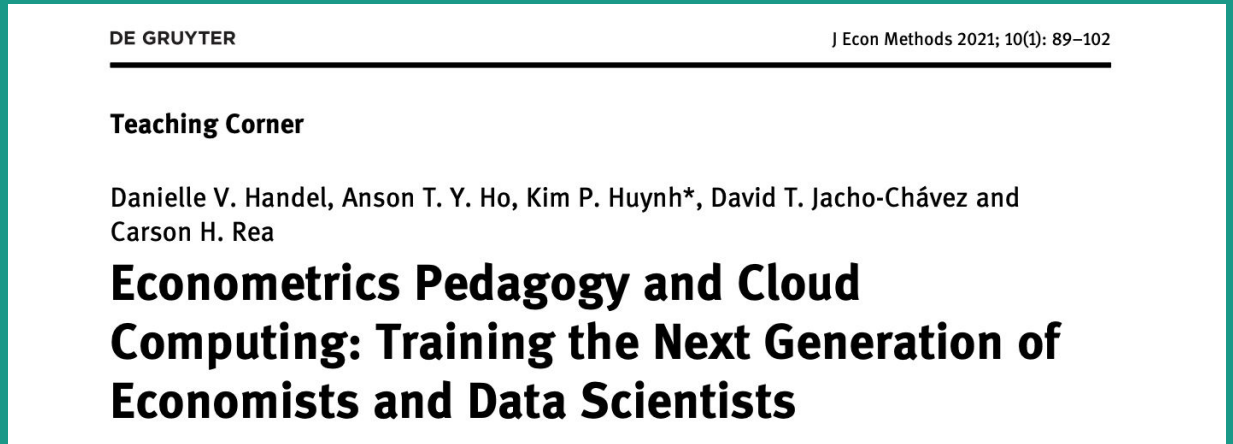
Toronto
Metropolitan
University

Stanford University

Cloud Computing Research Collaboration: An Application to Access to Financial Services

Danielle Handel, Anson Ho, Kim Huynh, David Jacho-Chávez, Carson Rea

Harnessing the power of on-demand cloud computing for collaborative research at the Bank of Canada



DE GRUYTER J Econ Methods 2021; 10(1): 89–102

Teaching Corner

Danielle V. Handel, Anson T. Y. Ho, Kim P. Huynh*, David T. Jacho-Chávez and Carson H. Rea

Econometrics Pedagogy and Cloud Computing: Training the Next Generation of Economists and Data Scientists

High Performance Computing (HPC) Resources at the Bank of Canada

- Edith2
  - On-premise HPC appliance
  - For use with sensitive data

- Digital Analytical Zone (DAZ)
  - Cloud computing platform
  - Supported by Microsoft Azure

Photo: Edith Whyte. ca 1966. Bank of Canada Archives (PC223-17) Credit: Unknown.

See: "Central bankers at the frontier: the state of the art in advanced analytics and AI" for a detailed description of this infrastructure

Cloud Computing for Research Collaboration

Advantages

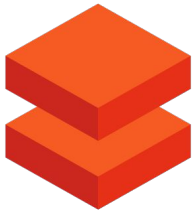
- On-demand
- Scalable
- Easy collaboration with external researchers
- Low startup costs

Considerations

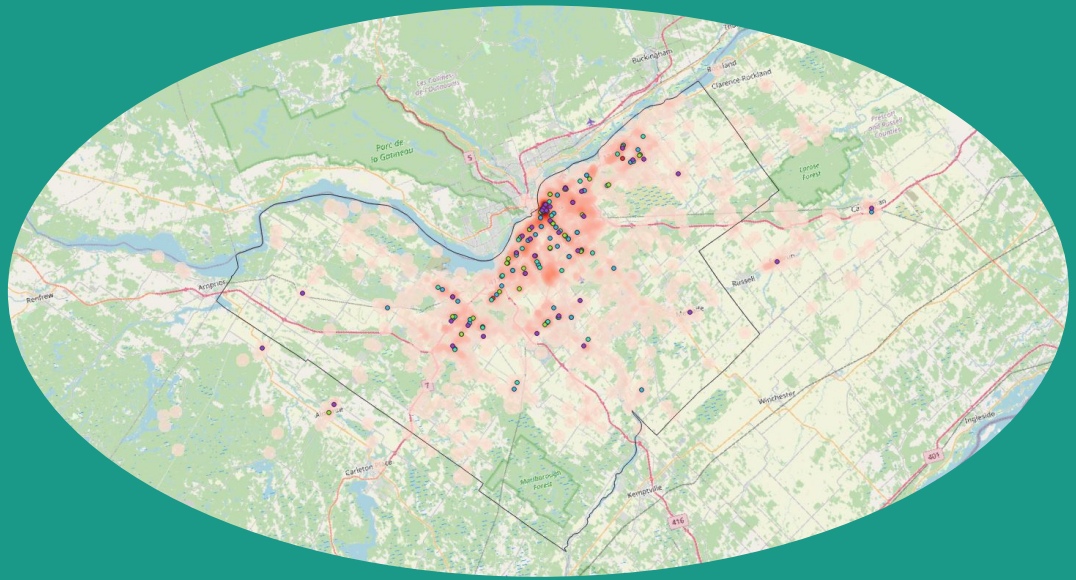
- Latency
- Training costs
- Budget restrictions

Microsoft Azure Databricks for Research Collaboration

- Jupyter-style notebook interface
- Fully managed clusters
- Machine learning tools
- Integrated version control



Empirical Use Case



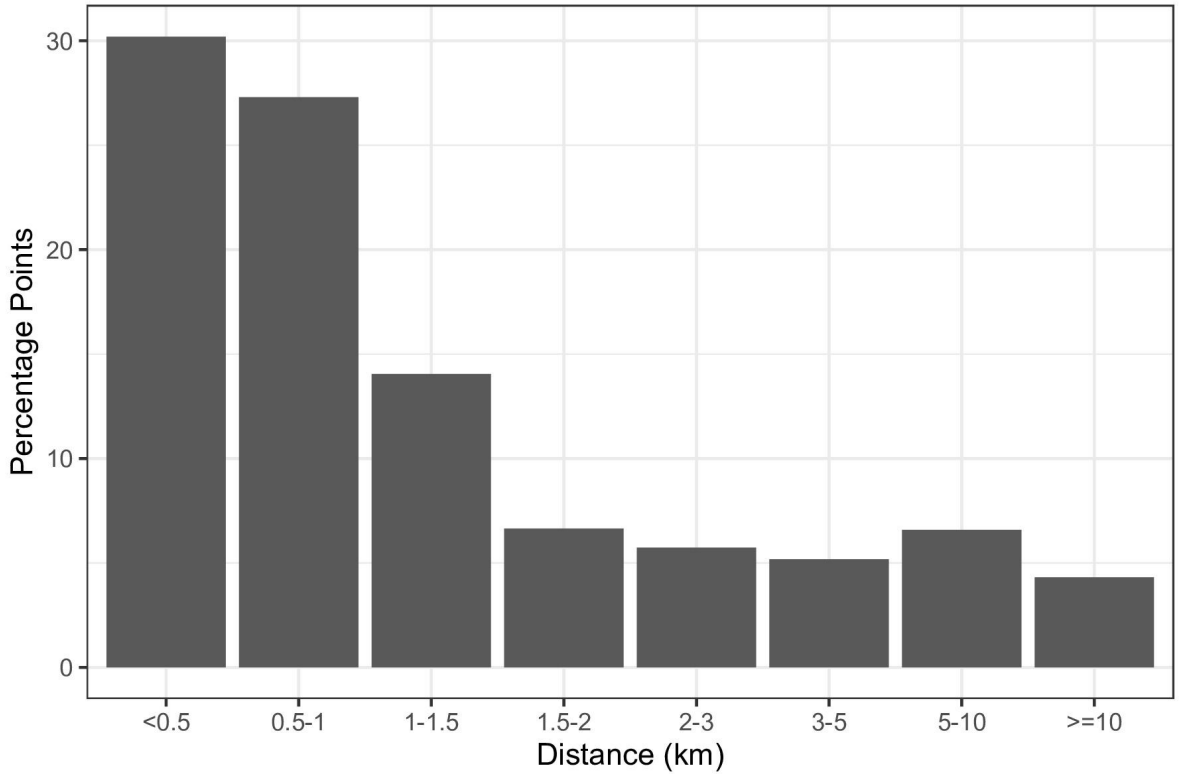
Access to Financial Services in Canada

Mapping consumers and their nearest bank branch

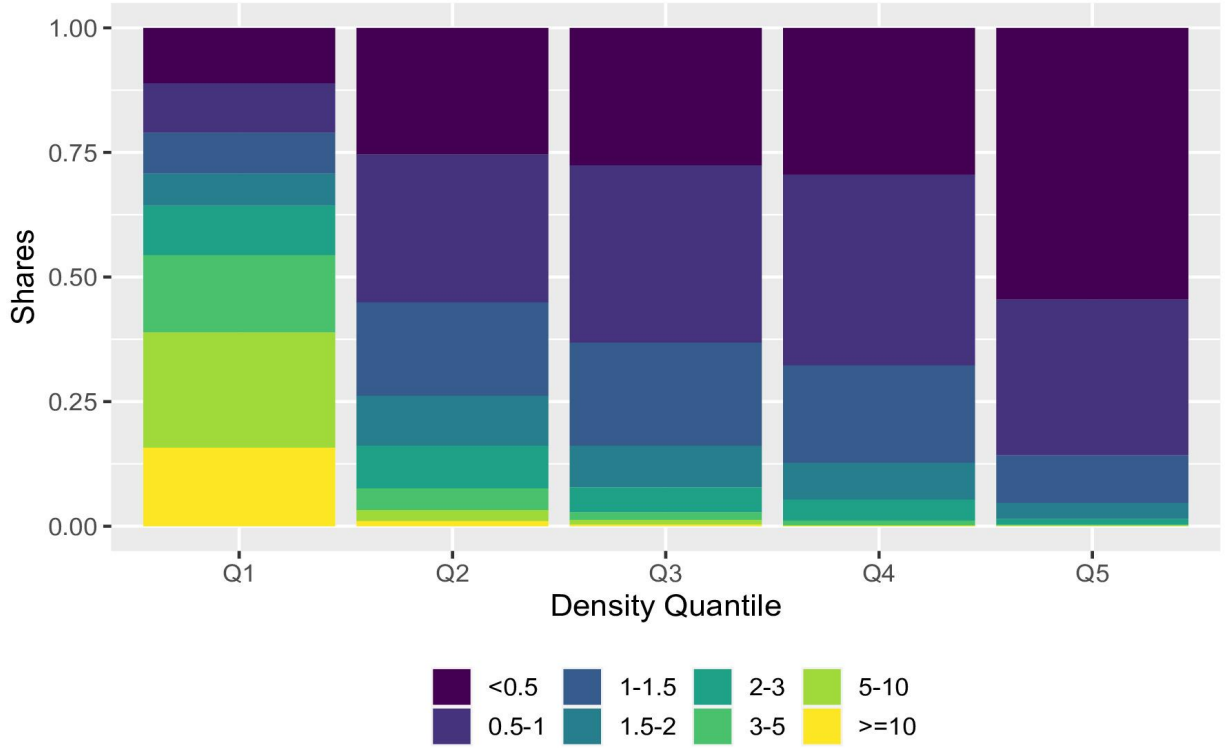
- Challenge: **24,000+** postal codes
- Computed straight line distances from population centroids to financial institutions
- Use Jupyter/PySpark docker image to manage computational needs

See: "A Spatial Model of Bank Branches in Canada" Staff Working Paper 2020-4 (English) by Heng Chen, Matthew Strathearn

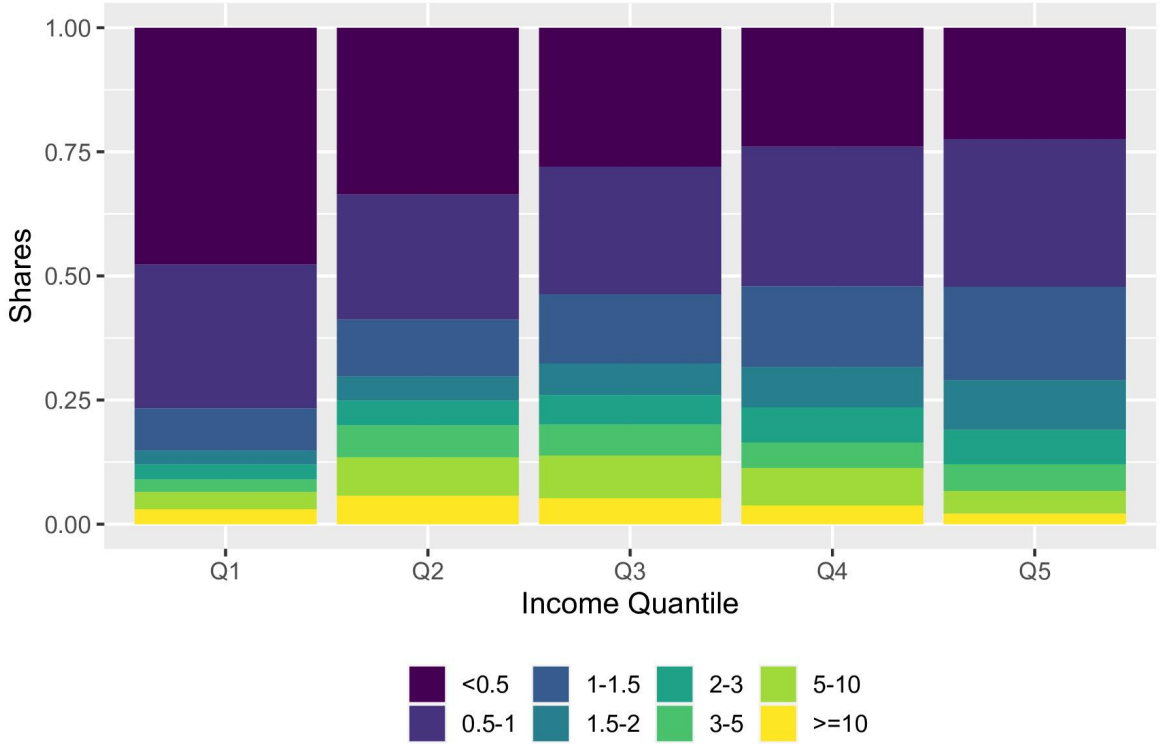
Over 50% of Canadians have access to a bank branch within 1 km



Banks are located in the most densely populated areas

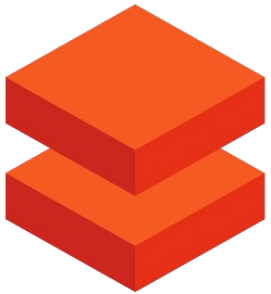


Bank branches are distributed across diverse income levels



The Added Value of Cloud Computing

- Custom, scalable resources
- Streamlined collaboration
- Low startup costs



Thanks/Merci

Danielle Handel
dvhandel@stanford.edu

Anson Ho
atyho@ryerson.ca

Kim Huynh
khuynh@bank-banque-canada.ca

David Jacho-Chávez
djachocha@emory.edu

Carson Rea
chrea@emory.edu