Rojas, Diego; Estrada, Juan; Huynh, Kim P.; Jacho-Chávez, David Tomás

Working Paper

Survival analysis of bank note circulation: Fitness, network structure and machine learning

Bank of Canada Staff Working Paper, No. 2020-33

Provided in Cooperation with:
Bank of Canada, Ottawa

Suggested Citation: Rojas, Diego; Estrada, Juan; Huynh, Kim P.; Jacho-Chávez, David Tomás (2020) : Survival analysis of bank note circulation: Fitness, network structure and machine learning, Bank of Canada Staff Working Paper, No. 2020-33, Bank of Canada, Ottawa, https://doi.org/10.34989/swp-2020-33

This Version is available at:
https://hdl.handle.net/10419/241199

BANK OF CANADA
BANQUE DU CANADA

Staff Working Paper/Document de travail du personnel — 2020-33

Last updated: August 18, 2020

Survival Analysis of Bank Note Circulation: Fitness, Network Structure and Machine Learning

by Diego Rojas,[1] Juan Estrada,[1] Kim P. Huynh[2] and David T. Jacho-Chávez[1]

[1]Department of Economics
Emory University, Atlanta, GA 30322-2240
drojasb@emory.edu; juan.jose.estrada.sosa@emory.edu; djachocha@emory.edu

[2]Currency Department
Bank of Canada, Ottawa, Ontario, Canada K1A 0G9
khuynh@bankofcanada.ca

ISSN 1701-9397

Acknowledgements

We thank Áureo de Paula, Jean-Frédéric Demers, David Drukker, Ben Fung, Ted Garanzotis, Harry J. Paarsch, Ramesh Paskarathas, Marcel Voia, and participants of the Advances in Econometrics conference on the "Econometrics of Networks" organized by the National Bank of Romania and the Faculty of Economic Sciences–Lucian Blaga University, Sibiu on May 16–17, 2019 for their various comments and suggestions on an earlier draft. We also acknowledge the efforts of Valerie Clermont, Ted Garanzotis, Mireille Lacroix, Andrew Marshall, Phil Riopelle, and Nathalie Swift and the use of the Bank of Canada's Digital Analytical Zone Microsoft Azure Cloud. We thank Meredith Fraser-Ohman for providing excellent editorial assistance. The views expressed in this paper are those of the authors. No responsibility for them should be attributed to the Bank of Canada.

Abstract

The efficient distribution of bank notes is a first-order responsibility of central banks. We study the distribution patterns of bank notes with an administrative dataset from the Bank of Canada's Currency Information Management Strategy. The single note inspection procedure generates a sample of 900 million bank notes in which we can trace the length of the stay of a banknote in the market. We define the duration of the bank note circulation cycle as beginning on the date the note is first shipped by the Bank of Canada to a financial institution and ending when it is returned to the Bank of Canada. In addition, we provide information regarding where the bank note is shipped and later received, as well as the physical fitness of the bank note upon return to the Bank of Canada's distribution centres. *K*-prototype clustering classifies bank notes into types. A hazard model estimates the duration in circulation of bank notes based on their clusters and characteristics. An adaptive elastic net provides an algorithm for dimension reduction. It is found that while the distribution of the duration is affected by fitness measures, those effects are negligible when compared with the influence exerted by the clusters related to bank note denominations.

Bank topics: Bank notes; Econometric and statistical methods; Payment clearing and settlement systems
JEL codes: E42, E51, C52, C65, C81

1 Introduction

Underlying and sometimes unseen network structures shape the observed outcomes across different markets. Human behaviours like crime, substance abuse, and educational achievement, among others, are significantly affected by the social network spanned by each individual (see, e.g., Jackson (2008) and references therein). More complex structures developed by individuals, such as financial systems, are also governed by network interactions. Exploring the network characteristics that affect an individual's behaviour provides important insights on the role that networks play in determining overall outcomes.

One particular market where the network structure is the cause as well as the consequence of human social interaction is the market of money circulation in a country. In Hobbes' (1996) Leviathan, there is a comparison of the economic structure of a nation as a human body, and bank notes are like the blood cells in the human circulatory system. The flow of bank notes among agents is a necessary condition for economic transactions to occur; therefore, the flow of bank notes literally fuels the economic engine of a nation. Following the circulatory system analogy, it is important to understand how this fuel that is being distributed among agents arises. This task is undertaken here in a novel way, going beyond the bloodstream and studying the cells themselves through the lens of big data analytics.

Our research uses the Bank of Canada's Currency Inventory Management Strategy (IMS) study of bank notes. IMS uses high-speed scanners to collect detailed images from each bank note in the Bank's possession. This results in a unique dataset containing records that track the events throughout the circulation cycle of every bank note issued by the Bank of Canada, as well as information about when the bank note was created, where it was shipped to, and when it was returned to the Bank. For this paper, we have a sample of over 900 million bank note unique scan records. The scan records span from August 2017 to July 2018. This dataset allows us to build networks at two levels: the region and the financial institution.[1] Specifically, for the latter, we examine the financial institution that receives the bank note from the Bank and the financial institution that returns it to the Bank.

We first present an exploratory analysis of this unique dataset addressing the bank note circulation patterns that can be identified. K-prototypes clustering is then implemented to classify bank notes together into clusters of bank note types. These clusters are then used as features when fitting a hazard model for the duration of circulation of bank notes in Canada. For this paper, the object of analysis is the duration of the first circulation cycle of the bank notes. The circulation cycle, or stay in the market, begins with the shipment of the note to the financial

[1] Due to confidentiality reasons, we do not name them.

institution and ends when the note is returned to the Bank of Canada. An elastic net algorithm is then used to select the best accelerated failure time (AFT) model for circulation duration. We find that while the distribution of the duration is affected by the scanned fitness measures, their effects are negligible compared with the influence exerted by the clusters associated with bank note denominations.

The paper is organized as follows: Section 2 describes how the data have been collected, their structure, and contained information. Section 3 provides insights about the recovered network structure implied by the data, while Section 4 explains how machine learning community detection (clustering) algorithms are used here to uncover bank note types. Section 5 provides the results from fitting a hazard model, while Section 6 concludes.

2 Single Note Inspection Data

2.1 Institutional Background

The Bank of Canada supplies financial institutions (FIs) with the bank notes they require to satisfy public demand through the country's Bank Note Distribution System (BNDS); see Bilkes (1997) for details.[2] At the centre of BNDS are the Bank of Canada's two cash processing centres, also known as Agency Operations Centres (AOCs), located in Montréal and Toronto.

The two AOCs are responsible for the distribution and secure storage of Canadian currency, mechanized note processing of financial institution deposits, and destruction of unfit note deposits. Under the distribution system agreement, these AOCs supply notes to FI Regional Distribution Centres (RDCs) located in 10 key cities across Canada, known as Regional Distribution Points (RDPs).

The inventory of bank notes is held at the RDCs and is distributed by the member FIs throughout their networks. In addition, the RDCs also supply bank notes to other FIs that are not part of the BNDS. The inventories are managed through an automated system used by the Bank of Canada and FIs, known as the Note Exchange System (NES), to facilitate the distribution, withdrawal and deposit of bank notes.

The NES maintains the record of inventories across all RDCs in Canada. The NES is used to identify and reallocate surplus inventories between RDCs, receive and fulfill withdrawal orders from FIs, and send instructions to the cash-in-transit (CIT) companies. In addition, the

[2]We thank Jean-Frédéric Demers (Director of the BNDS) for sharing his knowledge and the internal document of the BNDS which forms the basis of Section 2.1 Institutional Background.

NES provides a tool to forecast demand, and transmits the end-of-day balance to the Bank of Canada's general ledger.

The Bank of Canada maintains control over the flow and quality of bank notes by removing unfit notes from circulation and supplying new or fit notes as required. FIs ship their unfit notes to their RDCs, which are transferred by CIT companies and by air to the AOCs for processing. Paskarathas et al. (2017) provides a graphical representation of the BNDS system.

2.2 Structure of the IMS dataset

Recently, the Bank of Canada instituted single note inspection, which allows for a digital analysis of all bank notes in the IMS. The IMS is a unique dataset containing records that track the events throughout the duration of every note issued by the Bank. The collection of digital images for every note in the IMS has been part of the Bank of Canada's efforts to transition from a regularly implemented method of tracking bank notes based on sampling with human supervision to a more technologically advanced framework. Until late 2015, the Bank of Canada regularly sampled around 4 million notes from its inventory to try to assess the quality, or fitness, of batches of bank notes. The new system allows the Bank of Canada to track the movements of 22 million notes per month in a regular period, and up to 45 million notes per month in a peak period.

The new system uses high speed scanners to collect detailed images from each bank note. This allows the Bank of Canada to observe 22 fitness dimensions for each scanned note. From these recorded characteristics, the Bank of Canada can calculate statistics that summarize the wear and tear that each bank note has endured. A description of each of these 22 dimensions can be found in Table 1. The work of Paskarathas and Balodis (2019) uses the same set of fitness features and principal component analysis to try to establish associations among those note characteristics. The fitness status of note i is determined as follows: for each dimension d_{ik}, where $k = 1, \ldots, 22$, we observe the indicator function $\phi_{ik} = \mathbf{1}\{d_{ik} \geq l_k\}$. If $\phi_{ik} = 1$ we say that the note i is considered unfit on dimension k. For example, if any note has a "front stain" level greater than 8, we consider the note unfit on the dimension "front stain." Then, a note is considered unfit if $\max\{\phi_{i,1}, \ldots, \phi_{i,22}\} = 1$.

[Table 1 about here.]

Additionally, the data contain time stamps for: i) the shipment of the bank note from a Bank of Canada deposit centre; and ii) the receipt and scan of the bank note once it leaves circulation and is deposited back with the Bank of Canada by an FI. The data also contain geographical

information regarding the region of the deposit centre that shipped/scanned the bank note, and which FI required/deposited the note. These variables allow us to build a variable to describe what we call a "cycle," which is central to the analysis in this research. A cycle can be defined as the set of six elements:

1. Origin: the distribution centre from which each bank note was shipped.
2. Destination: the distribution centre where the bank note was deposited back.
3. Requesting FI.
4. Depositing FI.
5. Ship time stamp.
6. Scan time stamp.

Each bank note can have zero or more cycles associated with it. For example, a bank note that was shipped but has not been deposited back lacks three elements of the cycle, hence it has zero cycles associated with it. A bank note can also exhibit more than one cycle if it was shipped to the market and deposited back more than once. A graphical description of the life of a note is depicted in Figure 1.

In the cash industry, the *lifetime* of a note is defined as beginning when a note is first produced, at the moment of printing, and ending the moment it is destroyed. An important part of a bank note's life cycle is the time it spends in circulation (circulation cycle in Figure 1, bottom rectangle). Therefore, our analysis focuses only on the first time a bank note is put into circulation and when it is received back from circulation, i.e., its first circulation cycle.

[Figure 1 about here.]

The data in its pure form is a set of relational databases linked by a unique identifier: the bank note serial number. The construction of the cycle and its duration involves pairing records through the serial number on each of the tables on the relational dataset. After the matching process is performed, we retrieve the elements of a cycle associated with each bank note. Therefore, we can describe such cycles and build networks at three levels:

1. Region level: This is the less complex network where the edges are formed by the paths among 10 regions. These paths are constructed using data of the deposit centre region where the note was shipped, and the one where it was deposited back.

2. FI level: This has a medium level of complexity, and involves the financial institution that received the bank note from the Bank of Canada, and the financial institution that deposited back the bank note. For this case, a financial institution is assumed to be an independent unit in each of the regions that correspond to the deposit centre. For example, if Bank A has a presence in regions X and Y, then for the purpose of this network, this represents two nodes, Bank (A,Y) and Bank (A,X). This yields a network with 40 nodes.

3. Bank note as nodes: This level involves changing the unit observation to the bank note. In other words, each node in the network is given by a unique note. The edges are built around the paths they travelled, time frames, and other characteristics. This yields a network with around 6.7 million nodes.

Due to confidentiality requirements, only the region-level network is described in the next section. This paper analyzes solely the region level network. The analysis of more granular network structures is left for future research.

2.3 Sample description

The sample obtained for this analysis includes several data tables that provide a wide variety of information on individual notes. The shipment and scan data tables contain the most relevant information. The first contains around 600 million records of bank note shipments that range in time between July 2015 and April 2018; the second consists of nearly 300 million records of scanned notes corresponding to the period from August 2017 to July 2018. The first step in the analysis of this data is to eliminate possible duplicates in the data. We use the serial number of the note along with other characteristics to make sure we obtain unique records of notes up to a specific date.

The evolution of shipments and scans of the sample over time is depicted in Figure 2. The upper panels show only the evolution of shipments and deposits over time. The lower panels of Figure 2 show how scattered across time the shipments are by denomination in this sample. This is not surprising, since shipments are made in big lots. The deposits, which are the result of market factors, have a more consistent evolution; however, it is important to notice that the portion of notes whose denomination could not be identified is quite sizeable. In both samples, the 20-dollar notes have the greatest share. However, we cannot observe any shipment of 20-dollar notes in our sample after 2015. This is because the tracking of shipments of 20-dollar notes was not active after 2017.[3] We face a similar issue with the 50-dollar notes. Because the

[3]The tracking of shipments is still in a pilot phase; it is being rolled out for several denominations in different

shipments that are tracked are in a different period than that of the other denominations in the sample, we decided not to include the 20-dollar and the 50-dollar notes in our sample.

[Figure 2 about here.]

In addition to the previous information, the corresponding characteristics of the scan data were also obtained. There are roughly 900 million records for the information including all denominations. However, after the data cleaning process, the sample reduces to a total of 179 million records of shipments and a total of 165 million notes deposited. The final sample allows us to match all the events in which a bank note is involved, through its unique serial number and other relevant characteristics. In turn, this provides the opportunity to build the geographical paths where each bank note travelled. After matching, the data consist of around 6.7 million observations. We then filter for those notes that have more than one cycle. This is the main difference from the analysis performed by Paskarathas et al. (2017), where they allow for both right censoring and the possibility of multiple cycles. In contrast to this analysis, they focus on the total time it takes for a note to be considered unfit for circulation. Instead, we keep in the sample only those notes that have appeared once, at most. The particular group of notes that is left out is a small portion of data; around 98% of the notes are observed in the sample once, at most.

The 6.7 million matched notes provide a large amount of data that can be analyzed in several different ways. However, duration is the central feature of this research. Figure 3 shows the composition of the matched sample. First, the upper left panel shows the denomination prevalence. All three denominations are represented by a relevant number of notes in the final sample, and the most prevalent denomination is the CAN$10 note. This means that our analysis will have validity for any of those denominations. Next, on the upper right panel, we find the proportion of notes that have been deemed unfit by the scan process. The sample shows around 5% of the notes have been deemed unfit. While the ratio seems small, in practical terms this represents a sample of around 300,000 bank notes from which we can derive inference. The lower left panel shows the percentage of notes that were deposited back in each of the 10 regions. The largest shares correspond to regions 1, 5, and 7, which in turn corresponds to the regions that are able to draw more notes than the others. Since the Bank of Canada cannot control where these notes are deposited back, one reasonable explanation for this pattern is that it is a consequence of market transactions. Finally, the lower right panel shows the share of notes that were shipped to each region. The largest share of notes was sent to region 5. The notes are sent to the

periods of time.

market in each region according to the FIs' requirements. Interestingly, observe that for the case of region 1, the share of notes sent to that region is larger than the portion of notes that it deposits back. This result is a sign of movement of notes between regions that is not put into action by the Bank of Canada.

[Figure 3 about here.]

The elements shown in Figure 3 play key roles in the analysis regarding the geographical movements of the notes. The paths that the notes have travelled can be characterized by these four elements. This analysis constitutes a unique exercise to better understand the underpinnings of bank note circulation in the Canadian economy. Most importantly for this document, these elements will be crucial in understanding the determinants of the duration of circulation of a note in the market.

3 The Network and Spatial Patterns of Bank Notes

3.1 The Cycle Duration and Bank Note Fitness

The most notable feature when characterizing a bank note cycle is the time it takes to be completed, which we call the duration. Specifically, we focus on the time that passes between when a note is first shipped from the Bank of Canada and when it is first returned. We can analyze the distribution of the duration with respect to different groups of bank notes. For example, if we group bank notes by denomination, we can observe how their time in the market differs between groups. Additionally, we can group the bank notes by their fitness status. Figure 4 gives a sense of the distribution of duration when we consider these two dimensions.

[Figure 4 about here.]

One can draw several implications from Figure 4. First, the median duration of a bank note is shorter for small denominations (5- and 10-dollar notes) compared with the large denomination (the 100-dollar note). This result is true regardless of the fitness status. Second, unfit bank notes, at the median, regardless of their denomination, show similar or larger duration values—e.g., the unfit 5-dollar notes have around the same median as the fit 5-dollar notes; the unfit 10-dollar notes have a larger median than the unfit 10-dollar notes; and the fit and unfit 100-dollar notes also have around the same median. Finally, the location and shape of the distribution of duration seem to follow denomination-specific patterns, i.e., changes in the location and shape distribution of duration are less noticeable across fitness status.

The first finding is related to the nature of the usage of bank notes. Lower denominations (CAN$5 and CAN$10) are involved in a higher number of everyday transactions. Therefore, it is not surprising that the CAN$5 bank notes have a shorter stay in the market than the CAN$10 bank notes. Since there is a high number of transactions, we can conjecture that the CAN$5 bank note should be more likely to be deemed unfit than a CAN$10 note at any point in time. A similar statement could be made when comparing CAN$5 with CAN$100, or CAN$10 with CAN$100. However, the CAN$100 bank notes' duration distribution behaves differently than expected.

The CAN$100 bank notes show a duration distribution that is vastly more dispersed than that of other denominations. Also, the duration distribution of these notes is considerably shifted towards zero with respect to the other two high value denominations. The median duration of the CAN$100 bank notes is even lower than that of the CAN$5 bank notes. One possible explanation is the different logistics and management of CAN$100 bank notes. For example, CAN$100 bank notes are not available at automated teller machines (ATMs); consumers would need to obtain these notes directly from the physical bank branch. This different management could end up accelerating the life cycle of the CAN$100 bank notes beyond the factors that explain the behaviour of the other denominations. Jiang and Shao (2019) pose a possible explanation for differential velocity based on retailer behaviour. Retailers, due to operational costs, are more likely to require lower denomination bank notes to make change for their customers. These retailers would keep lower denominations in situ, and use the large notes to make large deposits with financial institutions.

The second finding is more aligned with what one would expect. The bank notes in the sample deemed fit have, at the median, a shorter stay on the market than those deemed unfit. This result is true for the CAN$5, CAN$10, and CAN$100 bank notes. The pattern holds beyond the median. To demonstrate this more clearly, Figure 5 shows the cumulative distribution of duration for each denomination and by fitness status. Notice that fit note duration dominates the unfit for CAN$5 and CAN$100. In other words, for every percentile in the distribution of duration, the duration of fit notes is shorter than unfit notes. For CAN$10 the difference is difficult to observe on the graph. The results show that every percentile difference is significant, with different signs depending on the percentile, but close to zero. Hence, we cannot reject the hypothesis that distributions of the two samples are equal. This difference could be significant due to the large sample that induces very small standard errors in the estimations. It is worth mentioning that the CAN$100 duration shape is different than other denominations. It has a more concave distribution, which suggests a different behaviour in the market.

[Figure 5 about here.]

The last finding is related to the role that denomination plays in determining the length of stay in circulation. A suitable hypothesis is that the main driver of changes in duration is the denomination of the bank note. While fitness does affect the duration distribution, these changes are relatively small when compared with the changes induced by denomination. For instance, Figure 4 clearly shows how every distribution across the fitness status changes only marginally, yet the change is dramatic across denominations. One possible explanation for this behaviour is related to the reasoning presented before for the first finding: the use of notes and the profile of the users change dramatically with denomination.

Finally, we take a geographic approach to this analysis and break down the duration by region and fitness status. The results are presented in Figure 6, which shows heterogeneous patterns of duration across regions. Region 1 shows the most distinct pattern: for this region the whole distribution is shifted towards zero and it has a higher dispersion in duration than any other region. This behaviour can be caused by a situation where the notes scanned in this region have a shorter stay in the market compared with every other region. Without disclosing the name of this particular region, we can assert that there is no compelling economic reason to believe that this region could have such a different behaviour than other nearby regions. The importance of understanding the paths travelled by the notes, hence the network they trace, becomes self-evident.

[Figure 6 about here.]

3.2 Money Circulation Network

The previous sections have focused on the description of the velocity of bank notes' circulation, quality, and denomination heterogeneity. An additional analysis can be performed regarding the circulation patterns through the network formed by regions. As described in Section 3, the network spanned by different actors involved in the distribution of bank notes allows us to analyze the flow of notes on the edges of the network. This circulation network is a directed graph composed of non-reciprocal relationships. Furthermore, the quantity of notes that are being sent and received across different regions may be totally asymmetric. It is important to mention that the adjacency matrix that serves as a base for the construction of these networks has a diagonal of non-zeros.

The first way to visualize this is through an actual adjacency/transition matrix. Figure 7 shows the share of notes that were sent between regions. The vertical axis lists each region i to

which notes were sent. The horizontal axis lists each region j from which notes were redeposited and scanned. Then, each cell (i, j) represents the share of notes that were sent to region i and were redeposited at region j. It is natural to observe larger shares along the main diagonal of the matrix. In other words, notes are more likely to travel between regions that are closer to each other. Additionally, notice that the main diagonal concentrates around 60% of all note flows. The implication then is that the majority of the flows follow a geographic sensible pattern, i.e., the notes stay in the region where they were shipped originally. However, notice that a sizeable portion of notes, around 24%, are actually sent from region 5 bounded to region 1. This movement, which escapes the proximity logic, represents a physical movement where the notes travelled through the economy between two places that are geographically far apart.

[Figure 7 about here.]

A second way to approach the same problem is to represent the flow of notes as a network. We can map the flow matrix we observed in Figure 7 into a graph by letting the set of nodes be the set regions, and a set of hyperedges be the connections between regions. For example, a note that travelled from region i to j is going to be an element of the hyperedge (i, j). Then we can characterize each hyperedge by aggregating outcomes on each of the notes that belong in it. For the sake of tractability, we define weights for the edges as follows: the percentage of notes that one node is transferring to another as a proportion of the total quantity of notes transferred in the network. Figures 8 and 9 depict the circulation patterns across the 10 distribution centres that are observed in the sample. It is important to remember that these flows do not imply the direct shipment of notes between RDCs. Instead, they imply movements of notes caused by market transactions—notes that were originally sent to node i are being collected back in node k. The number associated with each node gives the code representing each region.[4] The size of the nodes is a representation of the in-/out-degree statistic of the node. For both graphs, the darker edges represent a higher volume of transactions between nodes. In Figure 8, the bigger nodes are the regions receiving relatively more bank notes. Conversely, the bigger nodes in Figure 9 are the regions from where relatively more bank notes are moving away. For instance, the fact that region 1 in Figure 8 is bigger for the CAN$100 bank notes means that large quantities of CAN$100 bank notes are being scanned in that region after being shipped to regions like 6 or 7. The principal destination of the notes can be inferred by the darkness of the edge.

[Figure 8 about here.]

[4]The computations for network structures and statistics were performed using the `networkx`, and the graph by `PyGraphviz` modules for Python 3.7.

From the two graphs, interesting patterns appear. First, when analyzing the behaviour across all notes, regions 1 and 5 are the recipients of most bank notes. The pattern does not change when we break down the analysis by denomination. Notice that in the graphs for CAN$5 and CAN$10, we observe the presence of more edges. Furthermore, these edges depict several numbers of bipartite relations between region 5 and the rest of nodes. This relationship is a clear indication that this region has a high centrality in the network representing the flow of these bank notes. However, this relationship is not the case when we analyze the CAN$100. The CAN$100 bank notes have a clearly different behaviour pattern in comparison to other denominations. Regions 1 and 7 are consistently receiving CAN$100 denomination bank notes from all other regions. Most importantly, notice that such flows do not lead to important flows in the reverse direction. Combining these findings with the information coming from the shares of notes staying and moving across regions in the matrix of shares in Figure 7, we can argue that a large share of notes, mostly CAN$100, travel in one direction, towards region 1, and stay there until deposited. This means that at least one-quarter of the notes travel a large geographical distance, with no particularly compelling economic reason, towards a single point that seems to only attract notes and not send them back. Intuitively, this would lead to the implication that there is a large number of economic transactions where region 1 absorbs CAN$100 notes from all regions, but very few where the opposite happens.

When we examine the out-degree flows, notice that the most important region across all bank notes is region 5. The pattern remains the same across denominations. This suggests that region 5 is a particularly important economic centre, because it circulates most of the bank notes in the economy. Interestingly, the role of region 1 diminishes in terms of out-degree. Intuitively, this means that this node plays a small role when it comes to sending bank notes to other regions, specifically in the case of CAN$100. To explain these flows, the next section uses data-driven techniques in order to use the characteristics of each bank note on the hyperedges.

[Figure 9 about here.]

4 Banknote Clusters

This section explores the characteristics of bank notes in more depth. To do this, we impose some structure on a highly dimensional dataset, and in doing so we uncover patterns that may relate to the network flows we documented in Section 3. The network we observed consists of nodes, regions, edges, and flows of bank notes. Hence characterizing the bank notes is equivalent to characterizing the flows of bank notes among regions. By classifying bank notes by their

characteristics, we are intrinsically creating heuristic relations among them. These relations are expressed as the condition of belonging to a particular set of bank notes called a *cluster*. The main purpose of clustering bank notes is to explore the configuration of each cluster regarding its bank note characteristics. This empirical exercise allows us to gain insights into the way bank notes are performing in the economy, and provides crucial control information for the estimation of conditional hazard models in the next section.

We can sum up clustering in its most basic conception as minimizing the distance among nodes inside the cluster while maximizing the distance between clusters. The most popular clustering method is "K-means," (see, i.e., Jain (2010) and reference therein), which was discovered independently in several fields (Steinhaus (1956); Lloyd (1982); Ball and Hall (1965); MacQueen et al. (1967)). In its simplest form, this algorithm classifies observations using the distance of each observation to some K-means on the feature space. The algorithm first picks K observations from the data, and labels them as *centroids*. Then, one calculates the distance of every point in the dataset to each centroid. For each centroid there is a cluster; each observation is assigned to the closest centroid. From this classification we can obtain a mean squared error (MSE). Next, a new centroid is computed as the mean of the just-constructed clusters. The clustering repeats itself until the MSE converges. For the specific dataset we are using, we cannot directly use such a method. This is because a large number of variables in our dataset are categorical variables. The K-means method has only been proven to converge locally in the context of continuous variables. To take this possibility into account, a modification to the K-means algorithm is needed. There are several possibilities to tackle this task; see Szepannek (2018) and references therein. However, we use the strain of methods that follow the *K-prototypes* algorithm in Huang (1998), particularly an `R` implementation by Szepannek (2016). For completeness and following the description in Szepannek (2018), we now summarize the method. The dataset consists of n observations and a set of p random variables, x_i, $i = 1, \ldots, n$. To start the algorithm we assume there exist k prototype observations μ_j, $j = 1, 2, \ldots, k$. We can construct $c_{ij} \in C_{n\times k}$ a binary partition matrix where $\sum_{j=1}^{k} c_{ij} = 1$. Then we can choose the partition matrix that minimizes the objective function:

$$\min_{C_{n\times k}} \sum_{i=1}^{n} \sum_{j=1}^{k} c_{ij} \cdot d\left(x_i, \mu_j\right), \tag{4.1}$$

where the distance function $d(\cdot)$ is defined as:

$$d\left(x_i, \mu_j\right) = \sum_{m=1}^{q} (x_i^m - \mu_j^m)^2 + \lambda \sum_{m=q+1}^{p} \mathbb{I}(x_i^m \neq \mu_j^m). \tag{4.2}$$

The first term concerns the first q continuous variables with the square of the Euclidean distance. The second term concerns the $p - q$ categorical variables. The latter, the count of mismatches, is penalized by a parameter λ which assigns less or more importance to the categorical data. In our implementation, the parameter is chosen "optimally" based on the data according to the guidelines in Huang (1998) and Szepannek (2018).

The clustering would be indicative that notes have grouped, almost naturally, into well-defined groups according to their characteristics. Therefore, the choice of number of clusters becomes very important. In other words, we need a criterion to specify the number of clusters used in the algorithm. Following common practices in the literature, we performed the procedure using several numbers of clusters ranging from 1 to 15. We also computed several metrics of precision of the method, e.g., C-index, Gamma index, Dunn index, among others; see Charrad et al. (2014). Along with these criteria we also used the number of clusters that explained the most significant changes in the sum of squared distances. This method is known as the *elbow test*.

After computing the clustering scenarios, the tests coincide in two clusters as good choices for K. To perform K-means clustering, we have used our sample of matched notes. The matrix $\mathcal{X}$ considers 60 features in total. The matrix has been first standardized in order to improve the performance of the K-prototypes algorithm. The features include: fitness measures (see Table 1), origin of the bank note, and destination of the bank note, among others. The results for the clustering exercise are shown in Figure 10. The clusters are somewhat evenly sized, with 53.1% and 46.8% of the notes on clusters 1 and 2, respectively. This classification allows us to benefit from the rich structure of the data to form groups, and after classification it also allows us to analyze a selection of features in each cluster. In this sense, we have employed this technique as a sort of dimensional reduction technique.[5]

[Figure 10 about here.]

The high dimensionality of the data poses a challenge to try to characterize each cluster in a meaningful way. Our approach is to select the features that are most relevant to the purpose of this analysis. The selected features to analyze in each cluster are:

1. **Duration**: period of time, in days, between the moment the note is shipped from the Bank of Canada and the moment it is deposited back into the Bank of Canada.

[5]The procedure was implemented using the module `clustMixType` in `R`.

2. **Scanned region**: share of bank notes that were scanned in each destination. As previously discussed, the destination of the bank notes is one of the traits that has shown more interesting results in Sections 2 and 3.

3. **Denomination**: specifically, we examine the proportion of bank notes by denomination within each cluster, as the denomination structure is a main predictor of bank note duration.

4. **Year of scanning**: proportion of bank notes scanned in 2017 or 2018. These are the two years for which deposit information is available.

5. **Year of shipment**: proportion of bank notes shipped in 2017 or 2018. These are the two years for which deposit information is available.

6. **Fitness status**: proportion of bank notes deemed fit by the IMS data study after being deposited.

The features we have selected respond to two criteria. They must be conceptually compelling and have shown relevancy in the description of the data. The description of each cluster, based on the previously mentioned features, is shown in Table 2. One observes that the mean duration is very different for each cluster; coincidentally, the mean duration decreases with the label of the cluster. Dispersion of the duration measurement is lower within the first cluster. An additional interesting trait picked up by the cluster classification is the diversity of the scan location of the bank notes. There is a region that is more prevalent in each cluster. In the case of cluster 1, almost half of the notes were scanned in region 5. In cluster 2, almost 60% of the notes were scanned in region 1. In terms of denomination, the structure is different. Cluster 1 has a large proportion, around 80%, of CAN$10 notes. Meanwhile, cluster 2 has around 70% of CAN$100 notes. The remainders, in both clusters, are composed of CAN$5 notes. Regarding the scan year, cluster 1 is composed almost exclusively of bank notes scanned in 2018. Cluster 2 has a larger portion, around 30%, of notes that were scanned in 2017. The shipping of notes in the sample was almost exclusively done in 2017. Regarding the fitness status of the bank notes, we can see that cluster 1 has a larger prevalence of fit notes, around 97%, compared with 92% in cluster 2.

[Table 2 about here.]

Table 2 shows the following profiles:

- **Cluster 1 -**[10−**slower**]: longest stay in the market, relatively lower dispersion; primarily scanned in region 5 in 2018. Composed of a majority of CAN$10 bank notes. Shipped almost entirely in 2018. With a 97% share of fit banknotes.

- **Cluster 2-**[100−**faster**]: shorter stay in the market, relatively larger dispersion; primarily scanned in region 1. Composed of CAN$100 bank notes. Shipped almost entirely in 2018. With a 92% share of fit bank notes.

Notice that the network we are trying to characterize has, until now, a geographical structure. Each deposit centre or region is considered a node. Hence we can characterize each node according to the notes it receives. Also, consider that in the link between our network and the bank notes, up to this point, the bank notes can be interpreted as elements in the hyperedges between regions. We can think of the deposit centres and the flow of notes as macro features of the network, while the bank notes themselves are micro features that characterize the macro measures. In this sense, and regarding the patterns observed in Section 3, it is important to analyze whether the classification of the bank notes into clusters is uniform across regions, i.e., across nodes. This could inform us about the cluster label and characteristics of bank notes that are more prevalent in regions that have relatively higher/lower flow of bank notes in/out. We break down the bank notes according to where they were scanned and obtain the proportion of notes per cluster within each region. Results are shown in Figure 11.

[Figure 11 about here.]

In terms of bank note flow, one of the most interesting patterns mentioned before is the strong connectivity between regions 1 and 5. The in-degree graph in Section 3 showed us a pattern of bank note flow that is not clearly explained by economic activity between these regions. In general, while heterogeneity does exist, it is possible to distinguish two groups of regions: regions 2 through 9 are predominantly composed of notes in cluster 1, while the notes in regions 1 and 10 mostly belong to cluster 2. This finding, along with the network description, allow us to confirm that region 1 is something of an economic puzzle itself. Taking into account the profile we built for cluster 2, this analysis shows that even from a data-driven analysis, the particular behaviour of region 1—which mainly receives CAN$100—is not similar to any other region's behaviour. These insights come from a quick overview of the heterogeneity of the distributions of the clusters.

5 Hazard Model for Bank Notes

In this section, we analyze what factors influence a bank note's duration in circulation. The analysis in Section 3 showed how changes in the duration of stay in the market are correlated to the denomination of the bank note. [6] We start by plotting the survival curves of a bank note cycle.[7] In this context, the duration of a cycle of a bank note starts to be measured at the time of shipment and ends at the time of scan. We are only concerned with the bank notes in the sample, hence we do not take into account any right or left censoring.[8] Figure 12 displays the Kaplan-Meier (KM) estimator of the survival curves for each denomination.

[Figure 12 about here.]

The results are consistent with the findings in Section 2. The duration in circulation of the CAN$5 notes starts to decay after a period of around 100 days. Their failure rate steadily increases from that point onward. A similar case can be built around the KM estimate for CAN$100 notes. However, notice that in this case the curvature reverses, meaning the notes are deposited back faster in comparison to the CAN$5 bank notes. Interestingly, the most noticeable behaviour is that of the CAN$10 notes. These notes start to be deposited back after 200 days in circulation, twice as long as the CAN$5. This result is unexpected, since these two notes should have some degree of substitutability in the market. Another interesting feature is the very pronounced slope of the survival curve in comparison to the other denominations. We observe that the CAN$10 survival rate goes from one to zero in a tight period between 200 and 400 days after the original shipping date.

We can analyze these observed features in the framework of the AFT models, see, e.g., Kalbfleisch and Prentice (2002). This class of models allows for the duration to be "accelerated" by observable features of the units. In this case, we can observe how the duration changes with respect to the features of the note. Since we are assuming neither left nor right censoring, the specification can be estimated through the implementation of the ordinary least squares (OLS) estimator. This in contrast to other work such as Paskarathas et al. (2017), who observe right censoring and have to experiment with different model specifications and distributional assumptions. Our specification for this model is then:

$$\ln(t_i) = x_i'\beta + \varepsilon_i = f_i'\psi + k_i'\gamma + z_i'\phi + \varepsilon_i, \tag{5.1}$$

[6] Paskarathas et al. (2017) conducted a study of the life cycles of polymer versus cotton substrate.

[7] The computations were performed using the `lifelines` module in Python 3.7.

[8] Since we are analyzing only one-cycle-complete spell durations, we can ignore left and right censoring as per our sample description in Section 2.

where t_i represents the duration that bank note i stays in circulation, f_i is a vector containing i's recorded features, k_i represents a set of dummies from the cluster classification, z_i is a set of interactions of origin, destination, timestamps, and the features included in f_i, and ε_i represents the standard regression error. In total, the number of regressors in this model is 360.

There is an issue of near multicollinearity given the sparsity of the data and the high number of categorical variables included as features. This problem tends to inflate the coefficients of the model and shrinks its standard errors. To avoid these problems, we implement statistical learning regularization methods. These methods impose a penalization on the model's objective function so that the most relevant coeffcients can be selected. Popular methods within this class include ridge regression and lasso. The elastic net proposed by Zou et al. (2009) combined with double machine learning by Chernozhukov et al. (2018) is used here instead. This provides an automated model selection mechanism as well as automatic inference in the resulting model.

The elastic net method, proposed by Zou and Hastie (2005),[9] can perform better than lasso, or ridge, because it allows for grouped selection. For example, the lasso tends to select one variable from a group and ignore the others that are highly correlated with it. In contrast, elastic net regularization encourages grouping effects while stabilizing the L_1-regularization path. This approach tends to numerically break under conditions of near-multicollinearity. This is the case of our data. We then use a two-step procedure (see, e.g., Zou (2006)) that provides a more numerically stable approach to the selection of parameters using our data, i.e., Zou et al.'s (2009) version of the adaptive elastic net. For the case of a linear regression, following Zou et al. (2009), assuming the error is normally distributed with variance σ^2, it solves the problem:

$$\widehat{\boldsymbol{\beta}} = (1+\lambda_2/n)\left\{\arg\min_{\beta}\|\mathbf{ln}(\mathbf{t_i}) - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda_1^*\sum_{j=1}^{p}\widehat{w}_j\left|\beta_j + \lambda_2\|\boldsymbol{\beta}\|_2^2\right|\right\}, \tag{5.2}$$

where $|\cdots|$ and $\|\cdot\|$ represent the L_1– and L_2–norms, respectively, and $\widehat{w}_j$ comes from a first-step implementation of the elastic net estimator. Specifically, the values $\widehat{w}_j$ are given by $\widehat{w}_j = (|\hat{\beta}_j^\star|)^{-\gamma}$. The coefficients $\widehat{\beta}_j^\star$ are the coefficients of the elastic net estimation process in the first step.

The tuning parameters $\lambda = (\lambda_1^*, \lambda_2)'$ are the only ones left to be chosen. Specifically, λ_1^* is a regularization parameter that measures the "strength" of the L-1 penalization term in the objective function. When $\lambda_1^* = 0$, estimating the model is equivalent to estimating a ridge regression. The other parameter, λ_2, determines the "strength" of the L-2 penalization term in the objective function. If we set $\alpha = \lambda_2/(\lambda_1^* + \lambda_2)$, one can see how close we are from a ridge

[9]The model was implemented in the package `glmnet` in R.

or lasso regression, i.e., $\alpha = 1$ means that the resulting estimator is equivalent to using a lasso estimator; when $\alpha = 0$ the fitted model is equivalent to fitting a ridge regression model. The results use the λ that minimizes the MSE subject to $\alpha = 0.95$ in a 10-fold cross-validation on a grid of 100 λs.[10]

To obtain inference, we used a novel approach developed by Chernozhukov et al. (2018). The estimation commonly referred to as double machine learning uses Neyman-orthogonal moments/ scores in order to achieve consistent estimators of the linear model coefficients. This estimator grants asymptotic normality of the coefficients, hence it allows for the estimation of confidence intervals for some parameters of interest. This method is more suitable in comparison to the traditional way to use regularization methods, which report confidence intervals using a post-selection linear regression. However, the cost of such a convenient feature is that we can only make these inferences on a subset of the coefficients. For these exercises, and in view of previous findings, the variables we are interested in are: fitness status, denomination, and cluster label. This allows us to discern between the influence of quality, denomination, and the characteristics associated with the profiles we mentioned in the previous section. The double machine learning algorithm implemented closely follows the algorithm in Chernozhukov et al. (2018) and was coded up using the `R` package `glmnet`. The implementation of the adaptive elastic net estimator uses 10 folds each. Additionally, an OLS is estimated along with a "naive elastic net." The last one corresponds to fitting an elastic net model to select relevant variables, and then using OLS on the selected set to estimate standard errors.

The result of the specification of the adaptive elastic net procedure is a list of four coefficients along with their standard errors. The naive elastic net fits a log-normal linear model that takes into account only those variables selected by the method. The OLS includes all variables which are not multicollinear. In the original model, the changes were modelled using a comparison case in order to avoid perfect multicollinearity that stems from the use of categorical variables and aliasing. The base case considered for this exercise was a CAN$5 bank note scanned in the month of January in the year 2017, shipped in the month of January of the year 2015, travelled from region 1 to region 1, belonged to cluster 1, was considered fit, and had no damage in any of the fitness measures.

The results of this exercise can be found in Table 3. Robust standard errors are used to construct the z statistics and tests. All coefficients in the table are significant at a 99% level of confidence. The coefficients are transformed to represent the rate of acceleration using an exponential function. First, notice that the ADA-Enet implementation does show differences

[10]The value of α was chosen based on the speed of convergence of the resulting MSE.

with respect to the other two estimation methods. However, these differences are very small. This means that after we implemented a procedure to avoid near-multicollinearity, both OLS and Naive Enet are close enough to the more robust results. In this particular case, the size of the data could play a fundamental role in making the problems of near-multicolinearity shrink. Since the three methods show a relatively similar size and sign, we interpret the coefficients for one of the implementations.

The coefficients shown in Table 3 follow the findings from previous sections. Take, for example, the coefficient on the CAN$100; it significantly decreases the duration with respect to the base case scenario, CAN$5 notes. The opposite happens with the CAN$10 coefficient; in this case the duration is positively affected, growing almost 2% with respect to the CAN$5. These results differ from previous ones since they are able to partial out the effect of around 280 other variables through the implemented selection and projection methods. The results confirm that the CAN$100 denomination shows a pattern of higher velocity. The explanation for this particular phenomenon is unclear. On one hand, it could be that preferences to use larger denominations for larger transactions increases the velocity of the CAN$100. On the other hand, it could be that retailers are not willing to deal with large denominations. Something that could shed light on this matter would be to repeat the same analysis including CAN$20 and CAN$50 notes. However, the present sample does not allow for that.

Another interesting insight that comes from Table 3 is the sign of the fitness status category. Results suggest that compared with the base category—a note with no fitness problems—changing the status to unfit does not dramatically alter the duration of the note. It does, however, have a significant effect of 0.4% over duration. The dwindling size of the effect of this variable can mean that the actual time when a note is deemed unfit is almost irrelevant to its duration or that it is reflected in the cluster.

The opposite happens when considering the cluster dummies. Notice how the labelling of the clusters has a very clear effect on the bank note duration with respect to the base case scenario. The size of the coefficients is higher in comparison to those of the fitness measures. The base case was built on CAN$5 notes that were split evenly across clusters. Remember that cluster 2 was primarily composed of CAN$100 bank notes, and the stay in the market of CAN$100 bank notes was particularly shorter than other denominations, especially the higher denominations. Also, cluster 2 was the most prevalent cluster label in region 1, where the anomalous behaviour was present. Hence, the results shown in Table 3 are consistent with our previous findings.

[Table 3 about here.]

The analysis performed in this section corroborates the previous findings. The most impor-

tant determinant of note duration is the denomination of the bank note. Other characteristics, and their interactions, can play a role in determining other moments of the duration distribution; however, the location of the mean is affected mostly by the denomination. While these results are not causal, they imply that any hidden factors, if they exist, that affect the duration are most likely denomination-specific.

6 Conclusion

Recent survey evidence has highlighted the decreased usage of cash for point-of-sale transactions in Canada (see Henry et al. (2018))), while Engert et al. (2019) document that cash demand in Canada, measured as cash in circulation relative to GDP, has been stable for decades and has even increased in recent years. To understand the potential difference in transaction versus non-transaction demand for cash, we exploit the IMS network data to uncover bank note flows in the economy. One of the most important components of these flows is the physical exchange and movement of bank notes. Understanding the characteristics of such a network, and decomposing each component of its units, allows us to better understand the way economic transactions occur in an economy. Exploiting the uniqueness of the IMS data from the Bank of Canada, we first explore such a type of network. While several features were described, the one that took the central role in this research was the bank notes' length of the stay in the market, i.e., the duration in circulation.

Our analysis of the IMS data finds that one determinant of the duration of bank notes in the market is the group of fitness measures of the bank note, which are statistically (not necessarily economically) significant. We also find that the denomination structure of the bank note is economically and statistically significant, after we analyzed the network structure traced by the bank notes' travels. However, the nodes that behave as principal senders and receivers persist across all denominations. We compute the weighted in-degree and out-degree statistics for each node in this directed network to analyze the importance of each geographical region in Canada. In addition, we use the share of flows to analyze the importance of each edge of the network. We impose structure through K-prototype clustering on the data to classify the bank notes according to their features. Not surprisingly, even with a battery of confounding variables, the bank note denomination emerged as one of the key clustering determinants.

To obtain more compelling evidence, we analyze the distribution of duration through survival analysis. The unconditional analysis, with a Kaplan-Meier estimator of the survival curve, shows that there are clear, different, and denomination-specific survival curves. Next, we use an

accelerated failure time model with an elastic penalty to assess the main drivers of the duration. The results show that the fitness measures are not the main determinants of bank note duration but the cluster labels associated with bank note denominations are.

Our results are novel in that we find the duration in circulation of a bank note is related to denomination-specific clusters. This result implies that certain bank notes are involved in specific types of transactions that make certain denominations more or less likely to be deposited back to the Bank of Canada. Understanding denomination structure will help to understand payment choice; see Chen et al. (2019). Due to the non-causal nature of our results, we can only speculate for now; however, denomination-specific patterns seem to govern this variable. There are a variety of extensions that can be performed. For example, research could be undertaken that focuses on the spatial and temporal features of the data to understand if there are underlying unobserved factors that may confound the results. These are left for future research.

References

Ball, G. H., Hall, D. J., 1965. Isodata, a novel method of data analysis and pattern classification. Tech. rep., Stanford research inst Menlo Park CA.

Bilkes, G., 1997. The new bank note distribution system. Bank of Canada Review 1997 (Summer), 41–54.

Charrad, M., Ghazzali, N., Boiteau, V., Niknafs, A., 2014. NbClust: An R package for determining the relevant number of clusters in a data set. Journal of Statistical Software, Articles 61 (6), 1–36.
URL https://www.jstatsoft.org/v061/i06

Chen, H., Huynh, K. P., Shy, O., 2019. Cash versus card: Payment discontinuities and the burden of holding coins. Journal of Banking & Finance 99 (C), 192–201.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., Robins, J., 01 2018. Double/debiased machine learning for treatment and structural parameters. The Econometrics Journal 21 (1), C1–C68.

Engert, W., Fung, B., Segendorf, B., 2019. A Tale of Two Countries: Cash Demand in Canada and Sweden. Tech. rep.

Henry, C., Huynh, K., Welte, A., 2018. 2017 Methods-of-Payment Survey Report. Tech. rep.

Hobbes, T., 1996. Hobbes: Leviathan: Revised student edition. Cambridge Texts in the History of Political Thought. Cambridge University Press, Cambridge, MA.

Huang, Z., 1998. Extensions to the K-means algorithm for clustering large data sets with categorical values. Data mining and knowledge discovery 2 (3), 283–304.

Jackson, M. O., 2008. Social and Economic Networks. Princeton University Press, Princeton, NJ, USA.

Jain, A. K., 2010. Data clustering: 50 years beyond k-means. Pattern Recognition Letters 31 (8), 651 – 666, award winning papers from the 19th International Conference on Pattern Recognition (ICPR).

Jiang, J. H., Shao, E., 2019. The cash paradox. Review of Economic Dynamics.

Kalbfleisch, J. D., Prentice, R. L., 2002. The Statistical Analysis of Failure Time Data, 2nd Edition. John Wiley & Sons, New York, NY.

Lloyd, S., 1982. Least squares quantization in PCM. IEEE transactions on information theory 28 (2), 129–137.

MacQueen, J., et al., 1967. Some methods for classification and analysis of multivariate observations. In: Proceedings of the fifth Berkeley symposium on mathematical statistics and probability. Vol. 1. Oakland, CA, USA, pp. 281–297.

Paskarathas, R., Balodis, E., 2019. Big Banknote Data. Keesing Journal of Documents & Identity (February), 3–6.

Paskarathas, R., Graaskamp, L., Balodis, E., Garanzotis, T., 2017. Polymer versus paper substrate lifespan calculations: the case of Canada. In: Montoya, I. (Ed.), Banknote Management for Central Banks. Research and Markets, Ch. 10, pp. 163–189.

Steinhaus, H., 1956. Sur la division des corp materiels en parties. Bull. Acad. Polon. Sci 1 (804), 801.

Szepannek, G., 2016. clustmixtype: K-prototypes clustering for mixed variable-type data. R package version 0.1-16.
URL https://cran.r-project.org/package=clustMixType

Szepannek, G., 2018. clustmixtype: User-friendly clustering of mixed-type data in R. The R Journal 10 (2), 200–208.

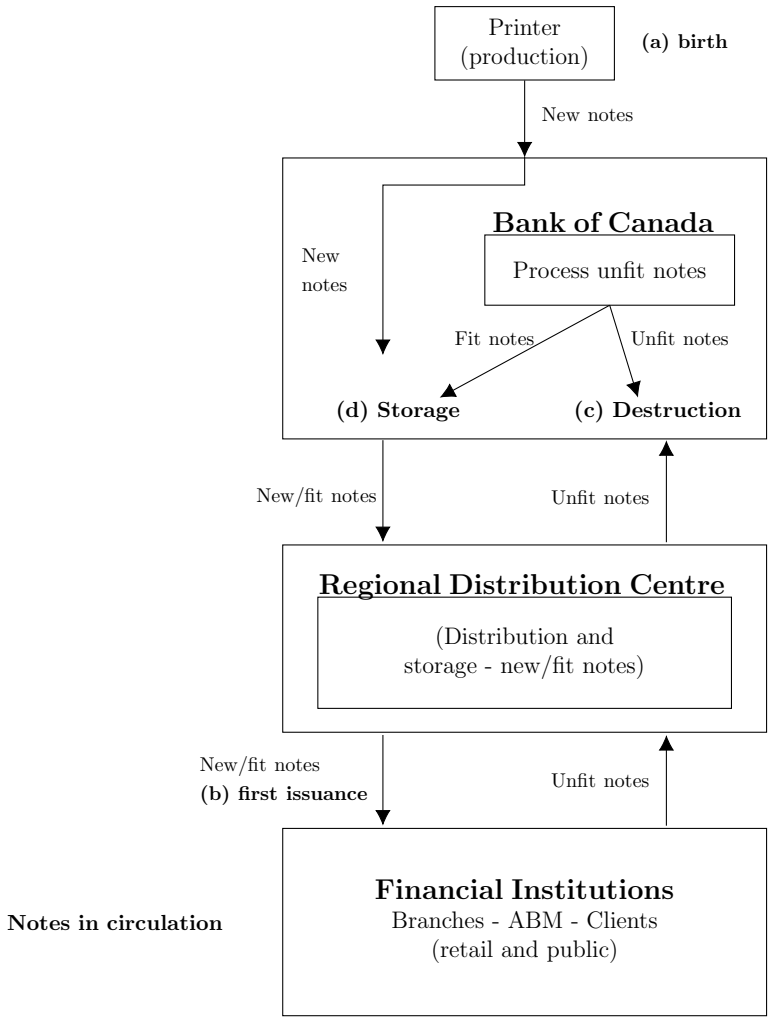
Zou, H., 2006. The adaptive lasso and its Oracle properties. Journal of the American statistical association 101 (476), 1418–1429.

Zou, H., Hastie, T., 2005. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society. Series B (Statistical Methodology) 67 (2), 301–320.

Zou, H., Zhang, H. H., et al., 2009. On the adaptive elastic-net with a diverging number of parameters. The Annals of Statistics 37 (4), 1733–1751.

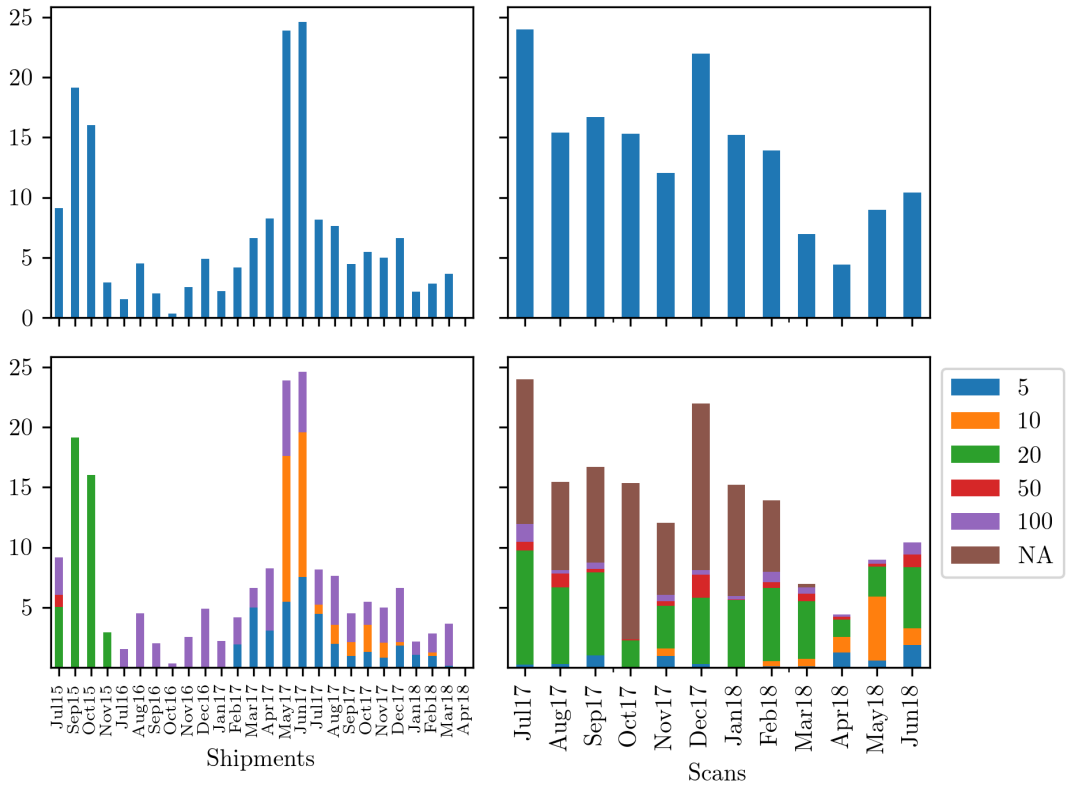
List of Figures

Figure 1: Bank Note Life Representation



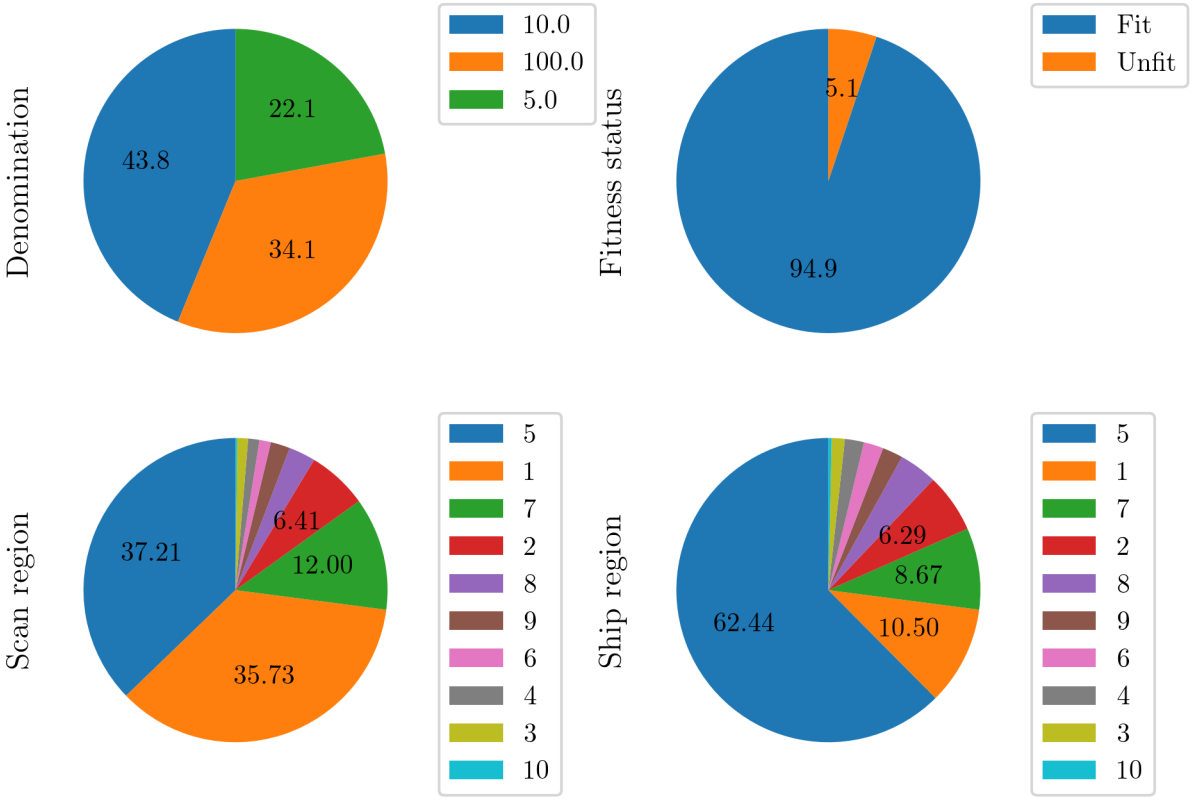
Note: A typical "cycle" any bank note would pass through while in circulation. This figure is reproduced from Paskarathas et al. (2017). The cycle can be tracked through each bank note's unique serial number. The life of a note is represented by the events (a)-(d). This paper focuses only on the first cycle of circulation; in other words, only the first time the note is shipped to financial institutions and returned to the Bank of Canada. Source: Paskarathas et al. (2017).

Figure 2: Sample Distribution over Time



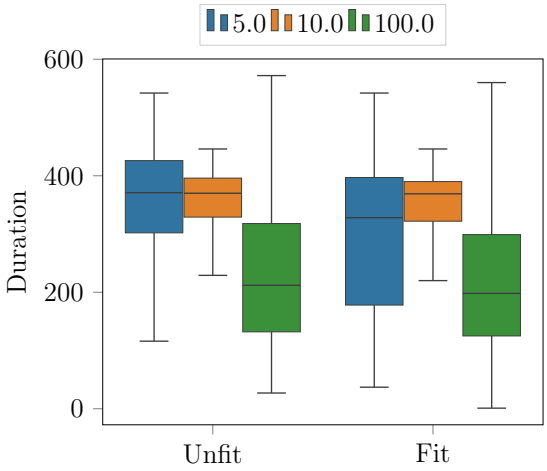
Note: The distribution of deposits and shipments behave differently over time. Both 20- and 50-dollar note shipments are more likely to occur before 2017 because the tracked shipments are in a different period.

Figure 3: Characteristics of the Matched Sample



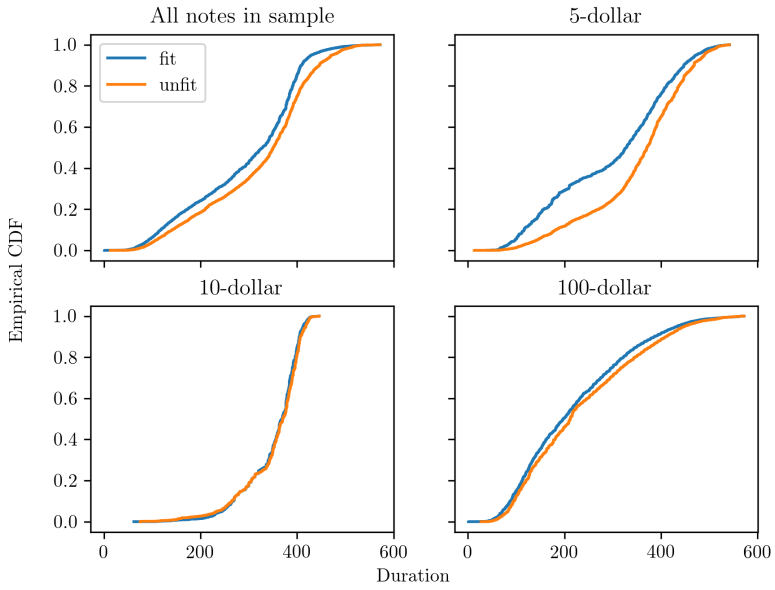
Note: Every denomination in the sample has sizeable share.

Figure 4: Bank Note Duration vs. Fitness Status



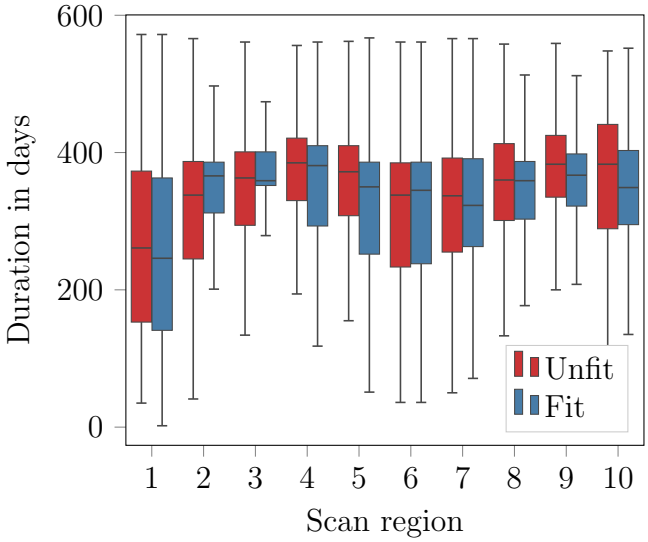
Note: Graph is based on the sample of 6.7 million bank notes matched from every source, with at least one cycle. Duration is measured in days, while "fit" and "unfit" follow the Bank of Canada's official definitions based on the recorded 22 dimensions of fitness. The boxplot shows the lower and upper quartile values of the data with a line at the median, and the whiskers extend from the box to show the range of the data.

Figure 5: Duration Distribution



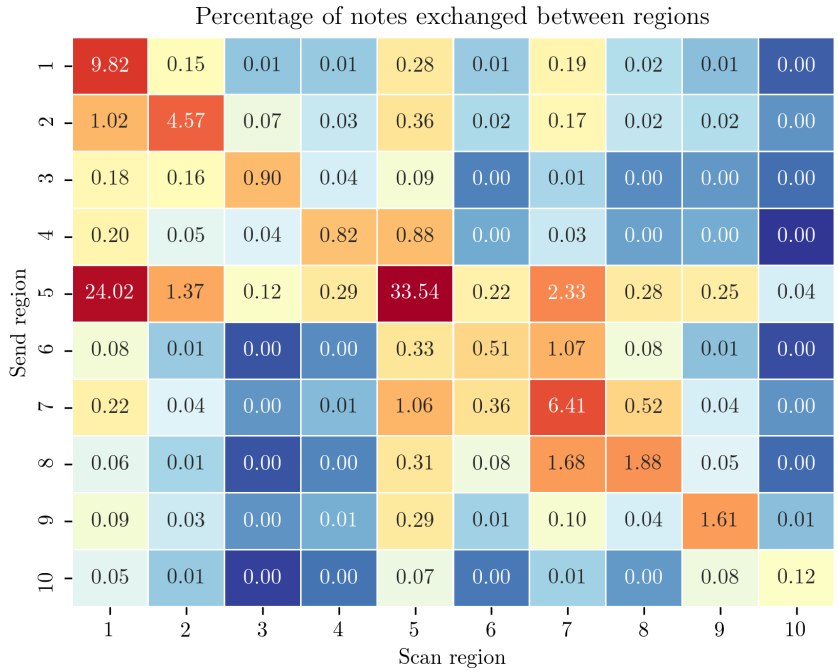
Note: Graph is based on the sample of 6.7 million bank notes matched from every source, with at least one cycle. Duration is measured in days, while "fit" and "unfit" follow the Bank of Canada's official definitions based on the recorded 22 dimensions of fitness.

Figure 6: Duration by Region and Fitness Status



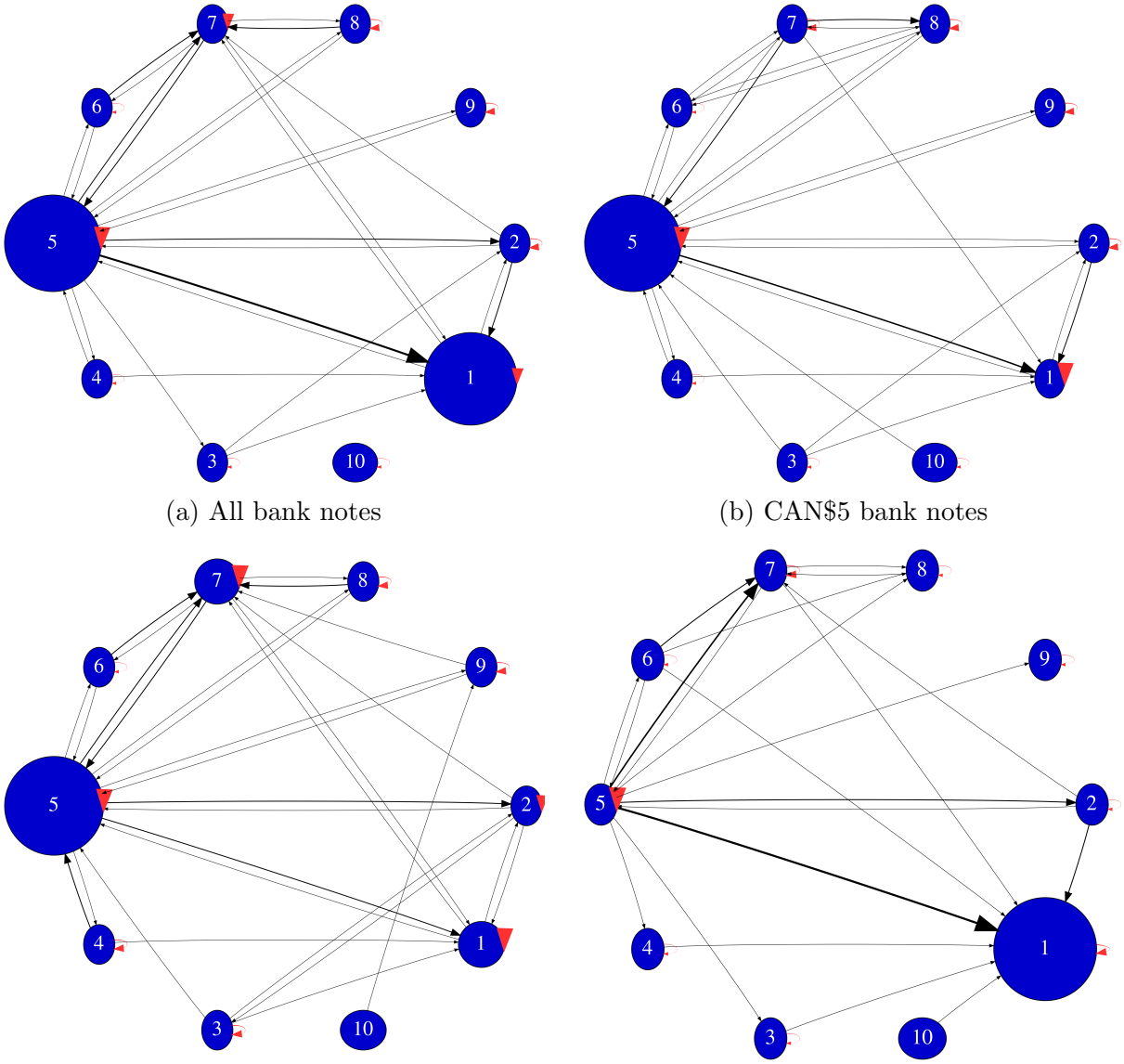
Note: Graph is based on the sample of 6.7 million bank notes matched from every source, with at least one cycle. Duration is measured in days while "fit" and "unfit" follow the Bank of Canada's official definitions based on the recorded 22 dimensions of fitness. The numbers from 1 to 10 in the horizontal axis are associated with the deposit centre region where the note was scanned.

Figure 7: Duration Distribution



Note: Graph is based on the sample of 6.7 million bank notes matched from every source, with at least one cycle. Each cell represents the portion of notes sent between regions. The larger the share, the closer the colour resembles deep red. The lower the share, the closer the colour resembles deep blue.

Figure 8: Directed Graphs for Bank Note Circulation Between Regions by Denomination (In-degree)



(a) All bank notes

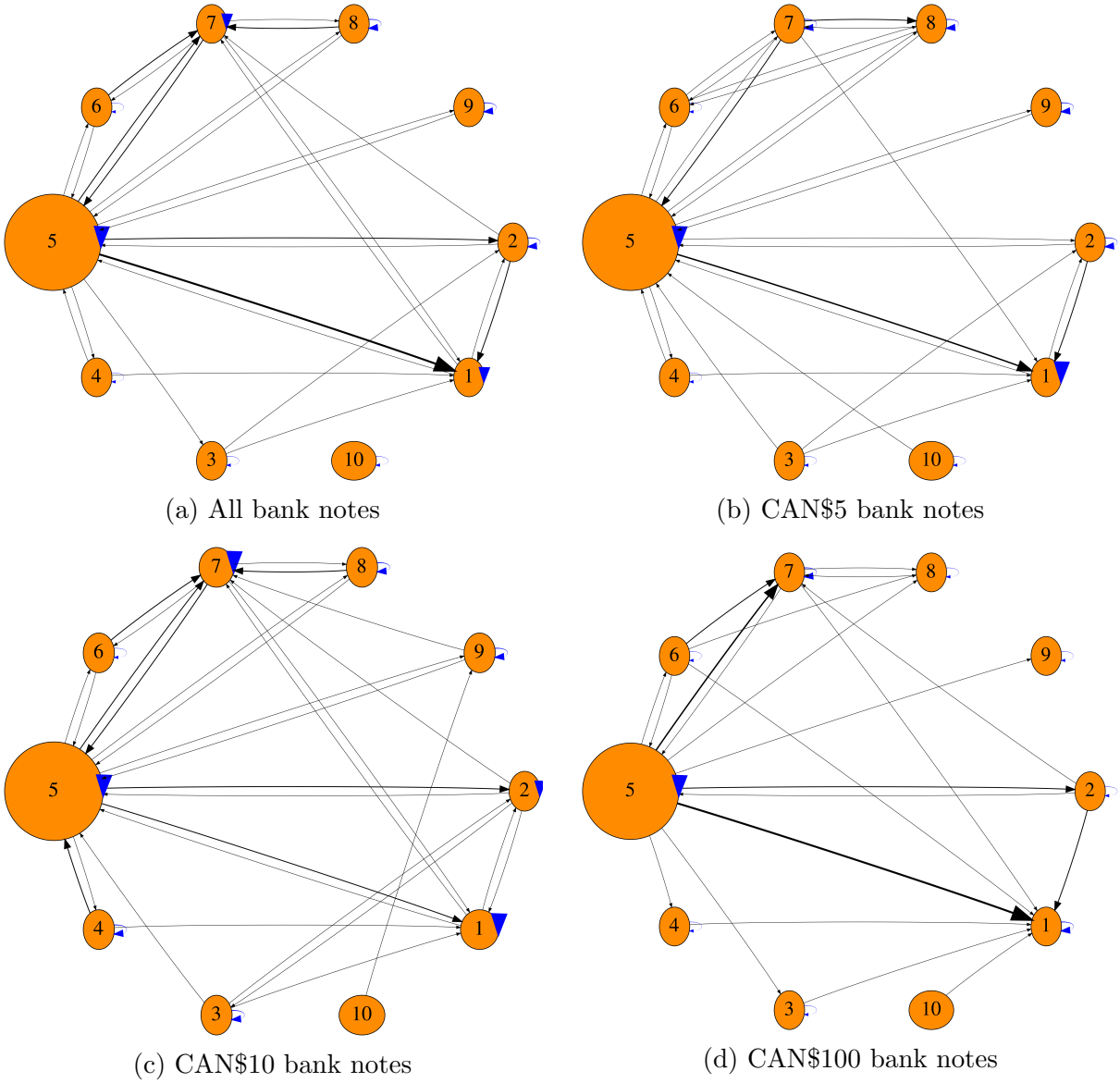
(b) CAN$5 bank notes

(c) CAN$10 bank notes

(d) CAN$100 bank notes

Note: The size of the nodes in each plot is given by a multiple of the "in-degree." The edges represent the share of all notes. They are coded in levels from thinnest to thickest following: less than 0.1% (transparent), 0.1% to 1%, 1% to 5%,5% to 10%, and greater than 10%.

Figure 9: Directed Graphs for Bank Note Circulation Between Regions by Denomination (Out-degree)



(a) All bank notes

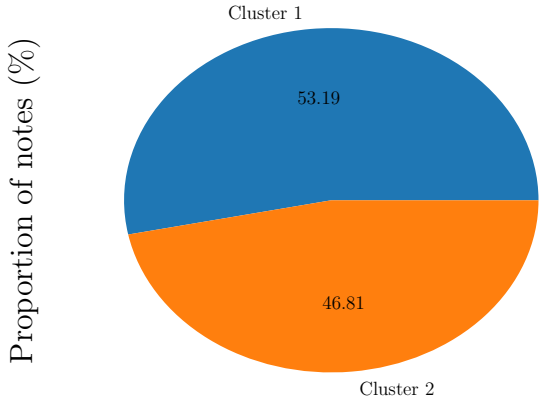
(b) CAN$5 bank notes

(c) CAN$10 bank notes

(d) CAN$100 bank notes

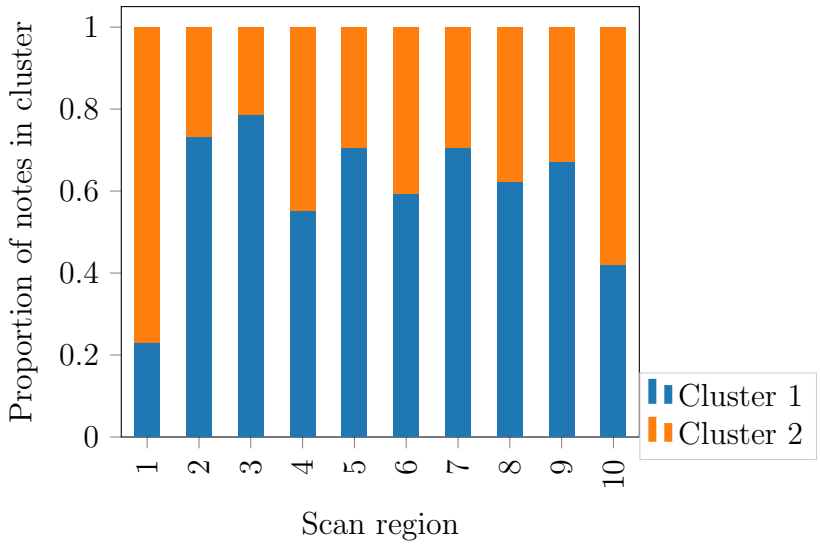
Note: The size of the nodes in each plot is given by a multiple of the "out-degree." The edges represent the share of all notes. They are coded in levels from thinnest to thickest following: less than 0.1% (transparent), 0.1% to 1%, 1% to 5%,5% to 10%, and greater than 10%.

Figure 10: Distribution of Bank Notes over Clusters



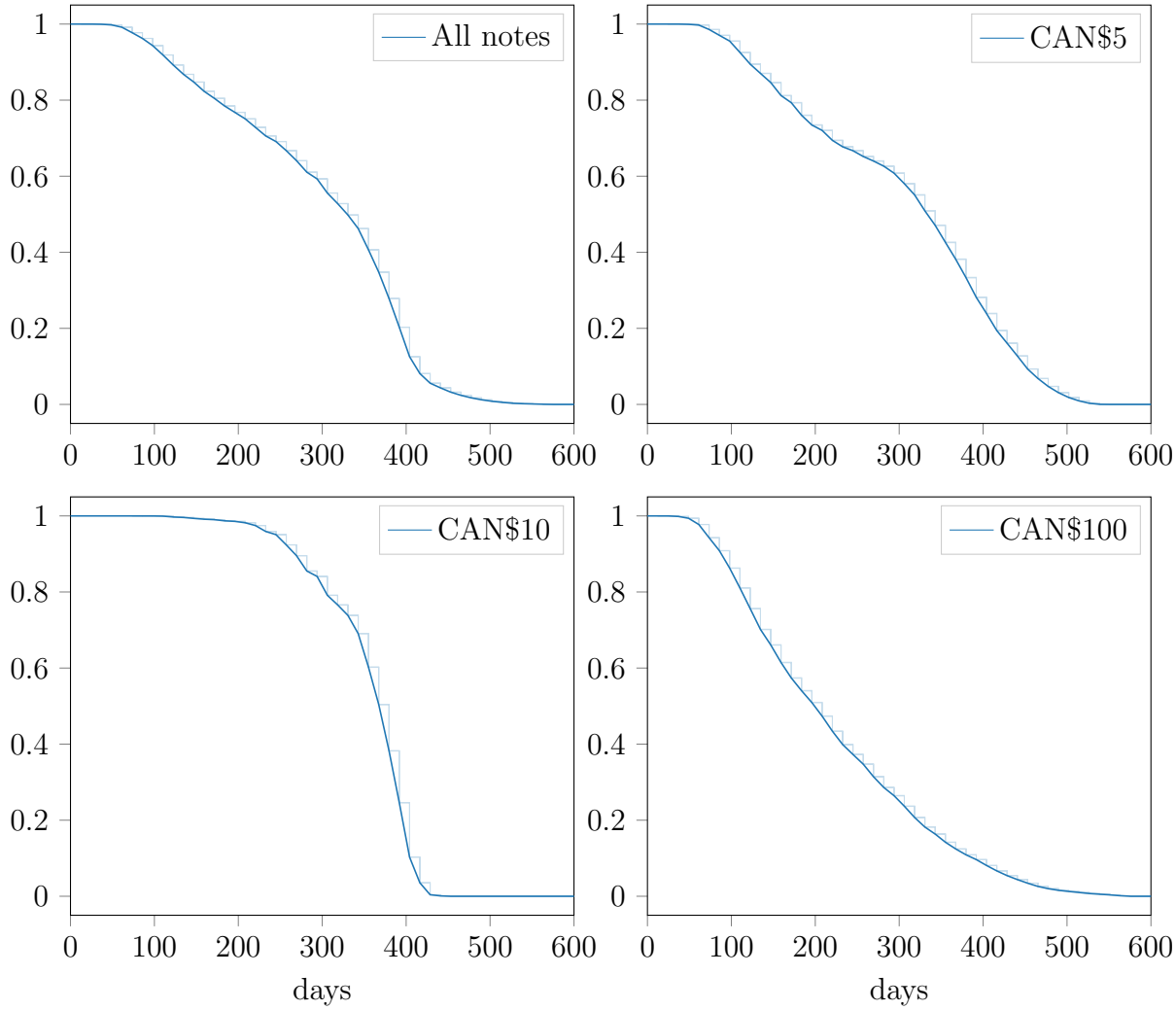
Note: This graph is the result of applying the K-means algorithm to the sample of 15 million matched bank notes that have completed one cycle at most. The clustering is done over the space of 60 features of the bank notes, including fitness measures, geographical, and spatial information.

Figure 11: Proportion of Bank Notes by Cluster Across Regions



Note: This graph breaks down the total number of matched bank notes by cluster and region where they were scanned. The heterogeneity in the proportions for each region shows how the flow of bank notes among regions differs greatly in terms of the types of transactions among regions and within each region.

Figure 12: Kaplan-Meier Estimates by Denomination



Note: This graph shows the Kaplan-Meier estimator of the survival curve for the duration of the bank notes' stay in the market. The estimate is based on the individual duration of each bank note. Each panel represents a subset of the sample according to the denomination of the bank note. The pattern that emerges is that the survival curve differs greatly, depending on the denomination. The notes with a faster rate of failure in their curves are the CAN$100 bank notes. The CAN$5 and the CAN$10 bank notes show a slope that is less steep.

List of Tables

Table 1: Fitness Dimensions Collected by Scanners

Category	**Parameter**	**Description**	**Threshold**
Holes	Hole	Sum of holes (sq mm)	3
Tears	Sum Open Tear	Sum of all open tears (sq mm)	4
	Max Open Tear	Longest open tear (mm)	2
	Sum Closed Tear	Sum of all closed tears (sq mm)	8
	Max Closed Tear	Longest open tear (mm)	8
Corner Faults	Missing Corner	Sum of all missing corners (sq mm)	15
	Folded Corner	Sum of all folded corners (sq mm)	70
Edge Faults	Missing Edge	Sum of missed edge (sq mm)	5
	Folded Edge	Sum of folded edge (sq mm)	5
Tape	Tape	Sum of tape area (sq mm)	100
Foil	Missing Foil	Missing foil area (sq mm)	40
	Sum Foil Scratch	Sum of foil scratch lengths (mm)	150
	Max Foil Scratch	Longest foil scratch length (mm)	20
Ink Wear	Ink Wear - Front	Ink wear front (levels 0-15)	8
	Ink Wear - Back	Ink wear back (levels 0-15)	8
Graffiti	Graffiti - Front	Graffiti level - total front (levels 0-15)	8
	Graffiti - Back	Graffiti level - total back (levels 0-15)	8
	Graffiti - Window	Graffiti level - clear window/foil (levels 0-15)	8
Stain	Stain - Front	Staining level front (levels 0-15)	8
	Stain - Back	Staining level back (levels 0-15)	8
Crease	Crease	Creasing/Crumpling level (levels 0-7)	5
Soil	Soil - Back	Soiling level back (levels 0-15)	8
	Soil - Front	Soiling level front (levels 0-15)	8

Note: Information is collected via high-speed scanners located in various Bank of Canada deposit centres once a bank note in circulation is deposited back in the Bank of Canada by a financial institution.

Table 2: Salient Features of Cluster Assignment

		Cluster	
		1	2
Duration	Mean	337.42	248.68
	Std. Deviation	78.03	125.31
Scan region	1	0.1550	0.5872
	2	0.0884	0.0365
	3	0.0171	0.0052
	4	0.0125	0.0114
	5	0.4931	0.2345
	6	0.0136	0.0106
	7	0.1596	0.0751
	8	0.0334	0.0228
	9	0.0260	0.0144
	10	0.0014	0.0022
Denomination	5.0	0.1482	0.3040
	10.0	0.8119	0.0132
	100.0	0.0398	0.6828
Scan year	2017	0.0648	0.2934
	2018	0.9352	0.7066
Shipment year	2017	0.9957	0.9424
	2018	0.0043	0.0576
Fitness status	Fit	0.0295	0.0761
	Unfit	0.9705	0.9239

Note: This table shows selected features of the clusters, obtained by performing K-means clustering on the 15 million matched bank notes in the sample. The patterns shown in this table allow us to construct profiles for the notes that belong to each cluster. The most salient traits are that the duration dramatically changes along with the denomination; also, the denomination composition almost exclusively segregates bank notes of one denomination to a specific cluster.

Table 3: Accelerated Failure Time: Bank Note Duration

	A-Enet		OLS		Enet	
	Coef.	A. F.	Coef.	A. F.	Coef.	A. F.
Intercept			5.8400	343.7793	5.8330	341.3813
			(0.0017)		(0.0015)	
CAN$10	0.0141	1.0142	0.0191	1.0193	0.0189	1.0191
	(0.0005)		(0.0005)		(0.0005)	
CAN$100	-0.0436	0.9573	-0.0428	0.9581	-0.04308	0.9579
	(0.0002)		(0.0002)		(0.0002)	
Unfit	-0.0043	0.9957	-0.0045	0.9955	-0.00455	0.9955
	(0.0003)		(0.0003)		(0.0003)	
Cluster 2	-0.0241	0.9762	-0.0247	0.9756	-0.02467	0.9757
	(0.0002)		(0.0002)		(0.0002)	

Note: This table shows the results of estimating an AFT model with adaptive elastic net (A-Enet), OLS, and naive elastic net (Enet), on the log-duration with respect to around 280 features and interactions, to select relevant regressors. To obtain the standard errors (SEs) enclosed in brackets, Chernozhukov et al.'s (2018) double debiased machine learning algorithm is performed. Coef. stands for the estimated coefficients, while Acc. Factor stands for their exponential. All coefficients shown in the table are significant at the 99% level. The base case considered for this exercise was a CAN$5 bank note scanned in the month of January in the year 2017, shipped in the month of January of the year 2015, travelled from region 1 to region 1, belonged to cluster 1, was considered fit, and had no damage in any of the fitness measures.