IFC-Bank of Italy Workshop on "Data science in central banking"

18-20 February 2025

From CPU to GPU: comparison of parallel computing techniques for numerical solutions[1]

Maxim Ralchenko, Oleksandr Shcherbakov and Kim P Huynh,
Bank of Canada

[1] This contribution was prepared for the workshop. The views expressed in this publication are those of the authors and do not necessarily represent the official views of the Committee, its members, or the BIS.

From CPU to GPU[1]

Comparison of parallel computing techniques for numerical solutions

Maxim Ralchenko, Oleksandr Shcherbakov, and Kim P. Huynh[2]

Abstract

This paper provides comparison of alternative parallelization techniques using central processing units (CPU) and graphics processing units (GPU) in application to a dynamic empirical model of consumer optimal cash consumption, holding and withdrawals. We document important optimization steps within programming code and compare OpenMPI versus CUDA implementations in terms of computation time, accuracy, and electricity consumption. To compare the practical efficacy we analyze what would it cost to implement these algorithms in a publicly-available cloud computing facilities.

Keywords: Demand for money, inventory models, heterogeneity, simulated maximum likelihood.

JEL classification: C15, C88, CE41, E42, L86.

[1] This paper is prepared for the BIS Irving Fisher Committee on Central Bank Statistics (IFC) 4th workshop on "Data Science in Central Banking," co-hosted with the Bank of Italy (BoI). We thank Giuseppe Bruno and Juri Marcucci for encouraging us to write this paper. We thank Nathalie Swift and the Analytical Environment Business Support (AEBS) team for their expert technical assistance on this project. This project is a collaboration between AEBS and Currency Economic Research and Analysis to modernize our structural econometric models that are useful for policy analysis. The content is not related to the economic outlook or to the direction of monetary policy. We acknowledge the use of the Bank of Canada's High-Performance Cluster EDITH3. The views expressed in this paper are solely those of the authors and do not necessarily represent those of the Bank's Governing Council.

[2] All authors are with the Bank of Canada. Contact email addresses are: mralchenko@bankofcanada.ca, ashcherbakov@bankofcanada.ca, and khuynh@bankofcanada.ca

1. Introduction

In economics, many interesting phenomena are described by complicated mathematical problems that are solved by many sophisticated decision-makers. For example, strategic consumer decisions of how much cash to consume and carry and how much and how often to withdraw can be formulated as a consumer dynamic programming problem with continuous dynamic controls. By matching model predictions to observed data we can estimate preference parameters which are useful in understanding determinants of behaviour of economic agents and are needed to accurately predict changes in the behaviour due to regulation or an exogenous shock.

When using parametric estimation techniques, to recover structural parameters of these models one needs to be able to solve them many times. When consumer heterogeneity becomes important, multiple solutions may be needed for multiple parameter values. For example, in simulated extremum estimators (simulated maximum likelihood or simulated method of moments) at least one solution is usually required per realization of a vector of random coefficients. As the dimensionality of random preferences becomes larger these models quickly become computationally infeasible.

Recent developments in empirical industrial organization, however, suggest faster ways of estimating rich preference heterogeneity by first presolving the underlying mathematical problem for a large number of candidate parameter values and then estimating a distribution of probability weights over agent types by matching model prediction to the observed data. These modern non-parametric methods to estimate preference heterogeneity were originally proposed by Ackerberg (2009)—more recently used by Malone et al. (2021), which is closest to our implementation, and McManus et al. (2022)—to our setting. Alternative approaches using semi-parametric mixtures estimators are discussed in Fox et al. (2011; 2016), Nevo et al. (2016). In this paper, we solve the problem more than 1.3 million times – one for each distinct vector of candidate parameter values. Consumer heterogeneity with respect to structural parameter values is combined with their heterogeneity along first the two moments of consumer income distribution and the variation in consumer cash withdrawal travel distance.

In this paper, we demonstrate how to implement these algorithms with general processing units (GPU) rather than a central processing unit (CPU). The idea of using GPUs rather than CPUs for dynamic programming for economic problem was discussed in Aldrich et al. (2011) and Aldrich (2014).

The rest of the paper is organized as follows. Section 2 outlines the economic problem and the solution algorithm. Section 3 describes various stages and layers of computer code optimization. Section 4 discusses performance of different hardware and software configurations employed in parallelization of the workload. Section 5 discusses the results.

2. Economic problem description

We study Canadian consumer cash management behaviour. Cash is valuable to consumers mainly due to its functionality as a medium of exchange and store of value. Our focus is on the use of cash as a medium of exchange, that is, consumer usage of cash as a payment instrument. Ability to use cash appears necessary for the store of value motive to realize. This next section is based and draws heavily from Huynh et al. (2025).

We assume that consumers have constant preference for cash and non-cash transactions at the Point of Sale (POS), can withdraw cash at an ATM or a bank branch, and can hold cash inventories. Cash withdrawal is costly. The cost is increasing in distance to the nearest cash access point and is independent of the withdrawal amount. Holding cash is also costly. The cost may reflect foregone interest, chances that cash is lost or stolen, need for safe storage for larger holdings, etc. Every period, consumers begin with a weakly positive level of cash holdings from the previous period and receive random draws defining their total current expenditures. We assume these draws are represented by independent identically distributed random draws from a normal distribution with mean and variance parameters to be estimated. After observing realizations of expenditure in the current period, consumers decide whether to withdraw cash and, if yes, how much, and how much cash to consume (i.e., spend in the current period). Let i denote a consumer type and let h_{it-1} be the level of cash holdings at the beginning of t. Let $s_{it} \stackrel{iid}{\sim} N\left(s_i, \sigma_i^2\right)$ denote current level of expenditures, and w_{it} and c_{it} denote levels of cash

withdrawal and cash consumption, respectively. We assume that consumer per-period reward function is defined as follows:

$$\begin{aligned} U_i\left(h_{it-1}, s_{it}, d_i; w_{it}, c_{it}\right) = & \alpha_i \ln\left(1+c_{it}\right)+\left(1-\alpha_i\right) \ln\left(1+s_{it}-c_{it}\right) \\ & -\gamma_i\left(h_{it-1}+w_{it}-c_{it}\right) \\ & -F_i d_i \times \mathbb{1}_{w_{it}>0} \end{aligned} \tag{1}$$

where α_i is a preference for cash transactions, γ_i is the cost of holding cash, d_i is a measure of distance to the nearest cash access point, and F_i is the fixed cost of withdrawal.

We assume that consumers maximize discounted value of utility flows over some time horizon $T \leq \infty$ by choosing the level of cash withdrawal $w_{it} \geq 0$ and cash consumption $c_{it} \geq 0$ in every period, i.e.,

$$\max_{(c_{i0}, w_{i0}), \ldots, (c_{iT}, w_{iT})} \sum_{t=0}^{T} \beta_i^t \mathbb{E} U_i\left(h_{it-1}, s_{it}, d_i; w_{it}, c_{it}\right) \tag{2}$$

$$\text{subject to } 0 \leq w_{it},\ 0 \leq c_{it} \leq \min\left\{h_{it-1}+w_{it}-c_{it}\right\} \tag{3}$$

where prior to making their choices consumers observe current period realization of total consumption s_{it}, and expectation is taken with respect to its future realizations at $t + 1$ onward, and the laws of motion for state variables are defined as

$$h_{it} = h_{it-1} - c_{it} + w_{it}, \tag{4}$$

$$s_{it} = s_i + \epsilon_{it}\sigma_i,\ \epsilon_{it} \stackrel{iid}{\sim} N(0,1). \tag{5}$$

Our goal is to solve problem (2) subject to constraints (3) for a given T. The solution is expressed as optimal cash withdrawal and usage policy for a consumer type i characterized by a vector of parameters $(\alpha_i, \gamma_i, F_i, \beta_i)$, observable distance state d_i, average spending level s_i and a known variance of income innovations, σ_i^2.

1. **Representative consumer model.** If we assume identical preferences for all consumer types i such that any consumer behaviour can be represented by the same vector of parameter values, estimation of the model parameters requires several hundreds of solutions (depending on how starting values are chosen). While there are material benefits from parallel computing even for the representative consumer specification, the real gain comes when we allow for semi-parametric or non-parametric consumer heterogeneity which is discussed next.

2. **Heterogeneous consumers.** While a representative consumer model is a useful starting point in our analysis, we find that observed consumer behaviour is likely to reflect significant variation in consumer preferences (and constraints) in population. A natural first step is to define consumer types using exogenous observable demographic characteristics such as age and income. With just two categorical values in each of the variables (low/high, young/old) we need to solve four problems of at least the same difficulty level as the representative consumer model discussed above. The `at least' part comes from the fact that non-linear search over much larger parameter vector requires significantly more function evaluations where each evaluation involves four solutions.

Conditioning on the observed demographics helps to improve model flexibility but it often cannot fully account for all the differences in consumer preferences particularly those determined by the unobserved consumer characteristics. Therefore, a richer non-parametric estimation techniques may be needed to accurately account for individual consumer decisions. This is particularly relevant when individual consumer-level data is available which is our case. In Huynh et al. (2025), we show that consumer observed behaviour suggests bimodality in consumer response to an increase in withdrawal costs. We find that some consumers can significantly reduce cash use, withdrawal level and the frequency of withdrawals, while others reduce cash consumption only sightly and increase withdrawal level while reducing the frequency. We find that while some of the behavioural differences are correlated with observable demographics, the non-parametric estimates reveal much reacher distribution of individual consumer responses than its semi-parametric analog does. Therefore, in what follows, we describe optimizations in the context of the non-parametric estimation method while results for a representative consumer specification and the one based on observable demographics are available in Huynh et al. (2025).

2.1 Solution algorithm

Our parametric assumptions do not allow for a closed-form solution of the dynamic programming problem (2). Obtaining a numeric solution for two dynamic controls is computationally challenging and we begin by re-formulating the problem in terms of a single control that uniquely identifies a withdrawal-consumption pair.

First thing to notice is that the effect of control variables (w_{it}, c_{it}) on the continuation value occurs via the change in cash holdings, i.e., $h_{it} = h_{it-1} + \Delta h_{it}$, where $\Delta h_{it} = w_{it} - c_{it}$. In other words, all withdrawal-consumption pairs such that $w_{it} - c_{it} = \Delta h_{it}$ have the same dynamic implications (result in the same continuation value).

Therefore, we can re-cast the dynamic programming problem (2) (with constraints (3) in terms of a single dynamic control Δh_{it} as follows,

$$\max_{\Delta h_{i0},\ldots,\Delta h_{iT}} \sum_{t=0}^{T} \beta^t \mathbb{E}\left[\tilde{U}_i\left(h_{it-1}, s_{it}, d_i; w_i^*(\Delta h_{it}), c_i^*(\Delta h_{it})\right)\right] \tag{6}$$

$$\text{subject to } \Delta h_{it} \geq \max\{-s_{it}, -h_{it-1}\}, \tag{7}$$

where $w_i^*(\Delta h_{it})$ and $c^*_i(\Delta h_{it})$ is a unique withdrawal-consumption pair that solves

$$\max_{(w,c)} \{\alpha \ln(1+c) + (1-\alpha)\ln(1+s_{it}-c) - F_i d_i \times \mathbb{1}_{w_{it}>0}\} \tag{8}$$

$$\text{subject to } w \geq 0,\ s_{it} \geq c \geq 0,\ w - c = \Delta h. \tag{9}$$

Lemma 1 in Huynh et al. (2025) establishes the uniqueness. Re-formulating the problem in terms of a change in cash inventories Δh allows reduces expensive numeric optimization to only one dimension.

To solve the maximization problem (6) numerically we define 30 grid points for each of the four state variables $S \equiv (h_{it-1}, s_i, \epsilon_{it}, d_i)$. We set $T = 183$ such that by making daily choices consumers solve optimal cash inventory problem with a half a year planning ahead horizon. The problem is solved using backwards induction. In the terminal period, consumers solve a simple static problem where exact solution is given by

$$(w_{iT}^*, c_{iT}^*) = \arg\max_{(w,c)} \{\alpha \ln(1+c) + (1-\alpha)\ln(1+s_{it}-c) - F_i d_i \times \mathbb{1}_{w>0}\} \tag{10}$$

$$\text{subject to constraints (9).}$$

As the solution is analytic, no non-linear optimization is involved at this stage. We assume that consumers do not incur any holding costs in the terminal period. Given the optimal solution we compute terminal period value function $v_{iT}(S)$ defined on 30^4 grid points in the state space. To compute value function at t = T − 1, for each point in the state space S we numerically search for optimal dynamic control Δh_{it}^*

$$\arg\max_{\Delta h}\left\{\tilde{U}_i\left(h_{it-1}, s_{it}, d_i; w_i^*(\Delta h), c_i^*(\Delta h)\right) + \beta \frac{\sum_{n=1}^{30} H\left(V_{t+1}(h_{it}, s_i, x_n, d_i)\right) f(x_n)}{\sum_{k=1}^{30} f(x_k)}\right\}, \tag{11}$$

where $H(\cdot)$ is a linear interpolation function approximating continuation value along the next period cash holding dimension h_{it} and the expectation over future realizations of income shocks is approximated by 30 quadrature draws x_n from a normal distribution with variance parameter, $\sigma_i^2(s)$, being a known (observed) function of average income level, and $f(x_n)$ are the weights given by the normal *pdf*.

2.1 Matching model predictions to data

Our individual-level data contains the following moments of cash withdrawal-consumption distribution on a daily basis: average cash consumption, average cash holdings, average cash withdrawal level conditional on withdrawing, average daily frequency of withdrawals, and the level of cash holdings that triggers a withdrawal.

Each consumer (an observation in the data) is characterized by a vector of state variable values, including average income level, variance of innovations to income in every period, distance to the nearest cash access point. In the data, consumers report their average cash usage, holding and withdrawal levels, withdrawal frequency as well as cash holdings at withdrawal.

For a given vector of parameter values, a solution to the consumer dynamic programming problem is represented by a dynamically optimal change in consumer cash holdings amount, $\Delta h_{it} = \Delta h(h_{it-1}, s_i, \epsilon_{it}, d_i)$ which, in turn, uniquely identifies a withdrawal-consumption pair in a given period, $(w_i^*(\Delta h_{it}), c^*_i(\Delta h_{it}))$. We are interested in the parameter values such that observed consumer behaviour is closely related to the model predictions.

We define "closedeness" of observed and predicted consumer behaviour in terms of simulated method of moments criterion function. More formally, on a daily level, let $\mathrm{E}[w_{it}|w_{it} > 0] \geq 0$ denote expected cash withdrawal level conditional on withdrawing, $0 \leq \Pr(w_{it} > 0) \leq 1$ denote expected daily frequency of cash withdrawals, $\mathrm{E}[h_{it}] \geq 0$ denote expected holding level, and $\mathrm{E}[c_{it}] \geq 0$ denote expected cash consumption. In the data, about half of observations for cash holding at withdrawal are missing. Therefore, we do not use this moment in estimation. The expectation is taken by averaging forward simulation outcomes over time. We define four structural error terms,

$$\varepsilon_{1i} \equiv \mathbb{E}[w_{it}|w_{it} > 0] - \widehat{\mathbb{E}[w_{it}|w_{it} > 0]}, \quad (12)$$

$$\varepsilon_{2i} \equiv \Pr(w_{it} > 0) - \widehat{\Pr(w_{it} > 0)}, \quad (13)$$

$$\varepsilon_{3i} \equiv \mathbb{E}[h_{it}] - \widehat{\mathbb{E}[h_{it}]}, \quad (14)$$

$$\varepsilon_{4i} \equiv \mathbb{E}[c_{it}] - \widehat{\mathbb{E}[c_{it}]}, \quad (15)$$

where terms without hats represent data sample moments and terms with hat denote model predictions. Let Z_{it} denote a vector of instrumental variables including a constant term, consumer income level and the distance to the nearest cash access point. We believe these variables are plausibly exogenous in our context. The method of moments parameters then become:

$$\theta = \arg\min_{\theta'} \{g_n(\theta')' \cdot W \cdot g_n(\theta')\} \quad (16)$$

where $g_n(\theta)$ is 12-by-1 sample moment conditions obtained by interacting each of the four structural errors defined in equations (12) through (15) with three instrumental variables and W is the weighting matrix. For the weighting matrix we use the optimal weighting matrix from the second stage of a representative consumer model. The representative consumer model is estimated using $(Z'Z)^{-1}$ (inner product of instrumental variable matrix) in the first and $W = (g_i(\theta^1)' g_i(\theta^1))^{-1}$ in the second stage (using first stage estimates θ^1).

Our version of non-parametric estimates of the distribution of consumer types is closest to McManus et al. (2022). In particular, having solutions for a large number of consumer types we estimate weights on each type such that the closest prediction obtains the weight of 1 and all other candidate types obtain zero weights. Matching analog would be $K-$nearest neighbour predictor with $K = 1$.

To implement this method we first need to obtain solutions for a large number of candidate types. We begin by solving the model for 1,300,000 consumer types, where each consumer type is represented by a vector of Halton numbers, $(\alpha_i, \gamma_i, F_i, \beta_i)$, $i = 1, ..., 1300000$. Halton numbers cover parameter space more evenly than pseudo-random numbers. As the original solution time using serial implementation was too slow (see Table 2), we considered parallelization solutions using OpenMPI with CPU and CUDA with GPU which are discussed next.

3. CPU and GPU computing

This section describes how we optimize the existing CPU code and then how we port this optimized code to GPU.

3.1 Parallelization of existing CPU code

The existing code can be parallelized through two main approaches:

1. Each combination of state variables (30^4 = 810,000 grid points), at each time step of the dynamic programming problem, see (11).
2. For multiple sets of observations, different sets of coefficients α_i, γ_i, F_i, and β_i to incorporate individual heterogeneity into the model.

The original CPU code was optimized with OpenMPI. Using the first optimization approach, there was no strong improvement in performance beyond 100 cores. The second approach does not require so much communication overhead and can be run on a greater number of cores, as the first approach requires synchronization of all grid points when moving to the next time step.

Based on findings from the GPU optimization exercise, we propagated a handful of optimizations back to th CPU code. These optimizations were found by identifying calculations that can be done at a certain point in advance and stored in memory, instead of repeating these calculations in an iterative process. The biggest of such improvements included:

1. Precalculating $\beta / \sum_{k=1}^{30} f(x_k)$ just once at the beginning as this term doesn't change throughout candidate solution values.
2. Precalculating the coefficients in the reward function, as most are consistent for all golden search iterations at a given time step
3. Precalculating the dot product $H(V(x)) \cdot f(x)$, as $f(x)$ is a constant, and $V(x)$ is calculated in the previous time step

As the code is numerically sensitive, precalculating and storing some coefficients yields very subtle differences in the final result, although different numerical techniques for the summation portion of the $V \cdot E_w$ dot product will give a great difference than the precalculation optimizations.

3.2 GPU porting

The existing C-OpenMPI codebase was converted to CUDA, which is an Nvidia proprietary API that permits the programmer to use the GPU for general-purpose programming on Nvidia devices. At a basic level, it permits the programmer to allocate memory on a device (GPU), copy memory from the host (CPU) to the device, and back, and invoke GPU kernels (snippets of code that execute on the device). For GPUs from AMD, who is the other major vendor, heterogeneous-computing Interface for Portability (HIP) is used. The HIP API is a C++ runtime and is a CUDA-like API. For both kinds of GPUs, OpenCL is another alternative for general-purpose GPU programming.

To write GPU kernels from scratch, usually a knowledge of C or C++ is required (alternatively, Fortran may be used, although this approach is much less common). The main learning curve to go onto the GPU for a programmer who is already experienced in those languages is implementing shared-memory parallelism. In this approach, multiple threads can access the same pool of shared memory. The difficulty with shared memory parallelism depends on the nature of the algorithm. For example, if some element of memory must be accessed many times, race conditions may easily occur, where two threads try to process some element of memory simultaneously. Alternatively, some algorithms, such as those classified as "embarrassingly parallel", are much simpler to implement, owing to the diminished risk of such conflicts.

The difference between shared memory on CPU versus a GPU computing is chiefly the nuances of GPU hardware. In a broad sense, GPUs were initially designed to calculate pixel values on computer monitors. These processing units are optimized at the hardware level to more efficiently process the arithmetic operations required to display graphics. Existing codes that handle embarrassingly parallel workloads and can operate in a SIMD (single-instruction, multiple-data) paradigm are particularly good candidates for optimization on a GPU; it is instructive to note that the "Hello World" example on a GPU is vector summation. Particularly useful resources for GPU computing on Nvidia include the CUDA C++ Programming Guide and the CUDA C++ Best Practices Guide. The latter publication is especially useful to optimize initial GPU code to account for points such as memory optimization (GPUs have many classes of memory that vary in size and read/write speed).

We exploit the fact that we can perform in parallel the calculations for each combination of state variables, at each time step. While the time-stepping procedure is by necessity serial, the calculations at each time step are done in parallel.

We initially used Nvidia A40 GPUs to run the converted code. The choice of GPU was based on existing availability, and is not the most ideal choice for this code, as it requires strong double precision performance. We subsequently acquired a Nvidia H100 GPU to further improve performance. Porting the code between the two GPUs is straightforward and can be done without changes; nevertheless, there are often optimizations that can be done to take advantage of the peculiarities of a specific GPU. In our case, a forward simulation, which kept the A40 fully busy, did not do so on the H100. An optimization was performed to execute multiple forward simulations simultaneously.

For this specific code, a GPU can be effectively used, because solving for each combination of state variables is an example of an embarrassingly parallel workload, as each combination at each time step is independent. Note that not every code is a good candidate for GPU optimization; examples of poorer candidates would be problems with fewer parallel tasks, or situations where each parallel tasks will tend to execute different processor instructions (termed divergence). Large-memory problems are not impossible to solve on a GPU, but there is more work that must be done to split the problem among multiple GPUs.

4. Performance and cost benchmarks

This section provides a speed comparison between CPU and GPU. In addition, we use the technical performance to provide a case study of how to potentially benchmark the cost of running the code on a publicly available cloud-based GPU.

4.1 Speed comparisons

Table 1 compares the execution time and speedups between a single CPU, an A40 GPU, and a H100 GPU. All of this hardware is available on the Bank of Canada Edith-3 High-Performance Cluster.

Code performance between CPU and GPU hardware

Table 1

	CPU	A40 GPU	H100 GPU
Execution time [minutes]	11.25	0.14	0.0054
Speed-up from serial	.	79	2077
Speed-up from H100	.	.	26

Notes: code performance is reported for a single observation in the data; CPU core is AMD EPYC 7543 32-core processor.

The speed-up factors can be thought of in terms of how many CPU cores are needed to replicate the GPU performance; in other words, 79 CPU cores would be needed to replicate the A40, but 2077 cores for a H100. This approach assumes little to no overhead between CPU cores for a distributed memory setup (i.e. MPI), which is a reasonable assumption for each CPU core handling a given combination of coefficients α_i, γ_i, F_i, and β_i.

The discrepancy between the A40 and H100 is due to the much improved double precision (fp64) performance on the H100; there is a theoretical ~50x performance difference for double precision between the two GPUs and a ~5x memory bandwidth improvement. The speedup factor of ~26x suggests that the code was mostly compute-bound as opposed to memorybound on the A40 (compute-bound codes spend more time doing calculations; memory-bound codes spend more time read and writing from GPU memory).

There are some power savings to be had with a GPU. Each compute node on our in-house cluster has 2 CPUs, and 64 cores in total. When all cores are running on a node, testing suggests that about 450W are consumed. To match the H100 performance, roughly 32 nodes are needed, which scales to 14.4 kW power usage. The H100 has a thermal design power (TDP) of at most 400W; if we assume that all cores of the node are operating (although, for the GPU code we only use 1), then we have at most 850W for the GPU code.

There is also a set-up cost question to consider. In a typical high-performance computing cluster environment, the cost to acquire ~2000 CPU cores is likely to be much greater than the cost of a data-centre HPC-optimized GPU. With consumer hardware, it is possible, as of 2025, to purchase a high-end (enthusiast-class) GPU that will have as much as 2.5x performance at double precision than the A40. It is therefore much more practical to purchase such a GPU, as opposed to a ~200-core CPU for this job.

4.2 Case study based on a normalization

We provide a case study to understand trade-offs. We focus on completion time or the actual time for a job to finish. To normalize all performance relative to the H100 GPU card which takes about 10 hours to finish one job. Based on the job time for CPU will ask how many cores would be required in a MPI job to match the 10 hours. We do similarly with a cloud-based A100 card. We choose the A100 as it is relatively affordable and widely available with many cloud-based services. It will provide a way for the reader to conduct sandbox or experiments without committing to large infrastructure investments.

Table 2 provides a comparison of the hardware required for CPU on premise (Edith3) and GPU cloud (A100) solution to match the wait time for the GPU on premise (H100). In addition, we estimate the cost of running the job on the GPU cloud based on GPU Azure.

Requirements to finish one job in 10 hours

Table 2

	CPU	GPU	GPUC
Hardware	AMD	H100	A100
Number of cores	2000	16896	55296
Clock speed	2.8 GHz	1.98 GHz	1.41 GHz
Memory	1 TB	94 GB	80 GB
Cloud cost per job	NA	NA	$200

Note: The GPU is on premise is the benchmark case as the job finishes in 10 actual hours. The CPU is on premise and GPUC is cloud hardware requirements are adjusted to ensure completion of the job in 10 hours. Info on GPU cloud is available here: azure.microsoft.com/en-us/pricing/details/machine-learning/

In our case the model lends itself to an embarrassingly parallel structure. In addition, the memory requirements are not so onerous since the dataset used is relatively small (about 10 megabytes). Aldrich (2014) provides a detailed discussion between using fewer CPU cores with faster clock speed versus much more GPU cores with slower clock speeds. From the Table 2 - CPU requires only 2000 cores at 2.8 Ghz relative to 16896 cores for H100 at 1.98 Ghz and 55296 cores for A100 at 1.41. Ghz.

A technical consideration is the memory requirements for CPU versus GPU. Aldrich (2014) makes a distinction between CPU memory versus GPU memory. CPU uses host memory while GPU uses device memory. Therefore, GPU instructions can only use data that are located in device memory. Therefore, it is important to have a fast interface between host and device.

The final calculation that we provide is the cost of running the job on GPU cloud. Based on our calculation it is about $200 per job. Hopefully, this cost calculation can help the reader to think about the tradeoffs in investing on on-premise solution versus GPU cloud solution. Due to the various complexities on-premise and total cost of ownership - we are unable to calculate the cost of the on premise solutions.

5. Discussion

As many interesting economic problems require computationally challenging solutions the role of parallelization techniques increases tremendously. Computational tasks that have been viewed as unfeasible using standard serial calculations, now become quick and efficient when solved using multiple processors at the same time.

We provide a concrete example based on policy-relevant research at the Bank of Canada. An important caveat is that we assume that the researcher is skilled in writing code in C and has the support of a dedicated high-performance computing specialist. We provide a brief guide on how to port the code from CPU to GPU.

We provide some calculations to help understand the tradeoffs inherent in implementing GPU versus CPU. We illustrated how calculation speed can be improved using GPU instead of CPU architecture as well as how these gains are increased with modern advancements in hardware development. We provide an example of case where GPU shines - embarrassingly parallel jobs which do not exceed the device memory requirements. Finally, we provide a back of the envelope calculation of how much it would cost to implement our example on publicly available cloud service.

6. References

Ackerberg, D. A. (2009). A new use of importance sampling to reduce computational burden in simulation estimation. QME, 7:343–376.

Aldrich, E. M. (2014). Chapter 10 - GPU computing in economics. In Schmedders, K. and

Judd, K. L., editors, Handbook of Computational Economics Vol. 3, volume 3 of Handbook of Computational Economics, pages 557–598. Elsevier.

Aldrich, E. M., Fern´andez-Villaverde, J., Ronald Gallant, A., and Rubio-Ram´ırez, J. F. (2011). Tapping the supercomputer under your desk: Solving dynamic equilibrium models with graphics processors. Journal of Economic Dynamics and Control, 35(3):386–393.

Fox, J. T., il Kim, K., and Yang, C. (2016). A simple nonparametric approach to estimating the distribution of random coefficients in structural models. Journal of Econometrics, 195(2):236–254.

Fox, J. T., Kim, K. I., Ryan, S. P., and Bajari, P. (2011). A simple estimator for the distribution of random coefficients. Quantitative Economics, 2(3):381–418.

Huynh, K. P., Shcherbakov, O., and Stenzel, A. (2025). Dynamic consumer cash inventory model. Technical report, work in progress.

Malone, J. B., Nevo, A., and Williams, J. W. (2021). The tragedy of the last mile: Economic solutions to congestion in broadband networks. Working Paper.

McManus, B., Nevo, A., Nolan, Z., and Williams, J. W. (2022). The steering incentives of gatekeepers in the telecommunications industry. Technical report.

Nevo, A., Turner, J. L., and Williams, J. W. (2016). Usage-based pricing and demand for residential broadband. Econometrica, 84(2):411–443.

Dynamic Consumer Cash Inventory Model

Kim P. Huynh
Bank of Canada

Alex Shcherbakov
Bank of Canada

André Stenzel
Bank of Canada

4th IFC Workshop on Data Science in Central Banking
Banca d'Italia
18-20 February 2025

Introduction

- rapid technological developments in the payment industry
  - → new payment instruments become available/are increasingly adopted
  - → consumers balance use of multiple instruments (cash, cards, mobile payments)
- cash usage at the point of sale (POS) has been declining over the past decade
  - → calls to abandon cash (Rogoff 2017), but retains market share (Henry et al. 2024)
  - → cash puzzle: demand for bank notes is steadily increasing (Engert et al. 2019)
    - ⇒ continued role for central banks to ensure accessibility of cash for entire populace
- infrastructure to access cash constantly evolving
  - → # of bank branches (predominant cash access point) in Canada peaked around 2013
  - → ongoing branch closure programs (e.g., Desjardins, Laurentian), reduction of up to 30% of branches

$\Longrightarrow$ Who are the winners & losers from these changes?

Contribution & Findings

- structural model of dynamic cash inventory management
  - → accounts for payment choice at point of sale: cash vs. non-cash
  - → estimation exploits multiple waves (2009, 2013, 2017) of detailed survey & diary data
  - → address role of changing infrastructure
- explicitly account for consumer heterogeneity
  - → representative consumer vs. fully flexible estimation approach: heterogeneity matters
  - → substantial heterogeneity even within demographic groups (old/young/rich/poor/urban/rural/...)
- consumer heterogeneity translates to bimodal response to changes in infrastructure
  - → substantially reduce cash use vs. infrequent but larger withdrawals & holdings
  - → younger and less affluent households forced to substitute away from cash despite preferences

Literature (non-exhaustive)

Baumol-Tobin type inventory models

→ Baumol (1952), Tobin (1956), Smith (1986), Whitesell (1989), Lippi and Secchi (2009) Alvarez and Lippi (2009; 2017), Chen et al. (2021) . . .

Technological Innovation & Payment Choice

→ Arango et al. (2015), Koulayev et al. (2016), Chen et al. (2017), Wakamori and Welte (2017), Huynh et al. (2020; 2022), Engert et al. (2024) . . .

Nexus of Cash Inventory Management & Payment Choice

→ Alvarez and Lippi (2017), Briglevics and Schuh (2020), Moracci (2022)

Cash Puzzle & Role of Cash

→ Rogoff (2017), Greene and Schuh (2017), Henry et al. (2018), Engert et al. (2019)

Heterogeneity Methods

→ Ackerberg (2009), Nevo et al. (2016), Malone et al. (2021), McManus et al. (2022)

Model: Overview (Parameters of Interest)

- Preferences (α): consumers want to use mix of cash & non-cash for transactions
  - $\rightarrow$ parameter on cash part of log-linearized Cobb-Douglas utility for payment choice at POS
  - $\rightarrow$ cash requires cash inventories comprised of cash holdings from previous period & withdrawals
- Costly withdrawals (F): interacted with distance (d) to cash access points
  - $\rightarrow$ distance measure d: taken from data, proxy for bank branch density in consumer's FSA
  - $\rightarrow$ cost of withdrawals depends on infrastructure (d) *and* preference/perception (F)
- Holding costs (γ): security, foregone interest, ...

▸ Details

$\Longrightarrow$ trade-off between frequent withdrawals and larger holdings to facilitate cash use at POS

One-Slide Model

4-slide version

- consumers use cash (c_{it}) or non-cash $(s_{it} - c_{it})$ to settle exogenous noisy consumption s_{it}
  - $\rightarrow$ cash uses inventory h_{it-1} carried over or costly withdrawal $w_{it} > 0$

$$\max_{(c_{it},w_{it})\forall t} \overbrace{\mathbb{E}\sum_{t=0}^{\infty}\beta^t U\left(h_{it-1}, s_{it}, d_i, w_{it}, c_{it}\right)}^{\text{discounted sum of flow utilities}} \quad \Longleftarrow \quad \text{reflects } \alpha, F, \gamma \text{ (log-linearized Cobb-Douglas)}$$

$$s.t. \ \underbrace{0 \leq c_{it} \leq h_{it-1} + w_{it}}_{0 \leq \text{cash use} \leq \text{cash} + \text{withdrawal}}, \quad \underbrace{c_{it} \leq s_{it}}_{\text{cash use} \leq \text{expenditure}}, \quad \underbrace{w_{it} \geq 0}_{\text{withdrawals nonnegative}}, \underbrace{h_{i,-1} = 0}_{\text{initial cash}}$$

- key: (c_{it}, w_{it}) impact future only via (implicit) choice of cash inventory h_{it}
  - $\rightarrow$ recast problem as choice of h_{it} implying conditionally optimal (c^*_{it}, w^*_{it})
  - $\rightarrow$ analytical characterization of conditional in-period choice
- approximate value function by backward induction using 6-months planning horizon
  - $\rightarrow$ $V(h_{it-1}, s_i, d_i, \epsilon_{it}) = \max_{h_{it} \geq \max\{0, h_{it-1} - s_{it}\}} \left\{\tilde{u}\left(h_{it}; h_{it-1}, s_{it}, d_i\right) + \beta \int V\left(h_{it}, s_i, d_i, \epsilon_{it+1}\right) dF_\epsilon\right\}$
  - $\rightarrow$ yields policy $h^*_{it}(h_{it-1}, s_i, d_i, \epsilon_{it})$ pinning down (w^*_{it}, c^*_{it})

One-Slide Model

4-slide version

- consumers use cash (c_{it}) or non-cash $(s_{it} - c_{it})$ to settle exogenous noisy consumption s_{it}
  - → cash uses inventory h_{it-1} carried over or costly withdrawal $w_{it} > 0$

$$\max_{(c_{it}, w_{it})\forall t} \overbrace{\mathbb{E} \sum_{t=0}^{\infty} \beta^t U\left(h_{it-1}, s_{it}, d_i, w_{it}, c_{it}\right)}^{\text{discounted sum of flow utilities}} \quad \Longleftarrow \quad \text{reflects } \alpha, F, \gamma \text{ (log-linearized Cobb-Douglas)}$$

$$s.t. \underbrace{0 \leq c_{it} \leq h_{it-1} + w_{it}}_{0 \leq \text{ cash use } \leq \text{ cash } + \text{ withdrawal}}, \underbrace{c_{it} \leq s_{it}}_{\text{cash use } \leq \text{ expenditure}}, \underbrace{w_{it} \geq 0}_{\text{withdrawals nonnegative}}, \underbrace{h_{i,-1} = 0}_{\text{initial cash}}$$

- key: (c_{it}, w_{it}) impact future only via (implicit) choice of cash inventory h_{it}
  - → recast problem as choice of h_{it} implying conditionally optimal (c^*_{it}, w^*_{it})
  - → analytical characterization of conditional in-period choice
- approximate value function by backward induction using 6-months planning horizon
  - → $V(h_{it-1}, s_i, d_i, \epsilon_{it}) = \max_{h_{it} \geq \max\{0, h_{it-1} - s_{it}\}} \left\{\tilde{u}\left(h_{it}; h_{it-1}, s_{it}, d_i\right) + \beta \int V\left(h_{it}, s_i, d_i, \epsilon_{it+1}\right) dF_\epsilon\right\}$
  - → yields policy $h^*_{it}(h_{it-1}, s_i, d_i, \epsilon_{it})$ pinning down (w^*_{it}, c^*_{it})

One-Slide Model

▸ 4-slide version

- consumers use cash (c_{it}) or non-cash $(s_{it} - c_{it})$ to settle exogenous noisy consumption s_{it}
  - → cash uses inventory h_{it-1} carried over or costly withdrawal $w_{it} > 0$

$$\max_{(c_{it}, w_{it})\forall t} \overbrace{\mathbb{E}\sum_{t=0}^{\infty} \beta^t U\left(h_{it-1}, s_{it}, d_i, w_{it}, c_{it}\right)}^{\text{discounted sum of flow utilities}} \Longleftarrow \text{reflects } \alpha, F, \gamma \text{ (log-linearized Cobb-Douglas)}$$

$$s.t. \underbrace{0 \le c_{it} \le h_{it-1} + w_{it}}_{0 \,\le\, \text{cash use} \,\le\, \text{cash + withdrawal}}, \quad \underbrace{c_{it} \le s_{it}}_{\text{cash use} \,\le\, \text{expenditure}}, \quad \underbrace{w_{it} \ge 0}_{\text{withdrawals nonnegative}}, \quad \underbrace{h_{i,-1} = 0}_{\text{initial cash}}$$

- key: (c_{it}, w_{it}) impact future only via (implicit) choice of cash inventory h_{it}
  - → recast problem as choice of h_{it} implying conditionally optimal (c^*_{it}, w^*_{it})
  - → analytical characterization of conditional in-period choice

- approximate value function by backward induction using 6-months planning horizon
  - → $V(h_{it-1}, s_i, d_i, \epsilon_{it}) = \max_{h_{it} \ge \max\{0, h_{it-1} - s_{it}\}} \left\{\tilde{u}\left(h_{it}; h_{it-1}, s_{it}, d_i\right) + \beta \int V\left(h_{it}, s_i, d_i, \epsilon_{it+1}\right) dF_\epsilon\right\}$
  - → yields policy $h^*_{it}(h_{it-1}, s_i, d_i, \epsilon_{it})$ pinning down (w^*_{it}, c^*_{it})

Data: Cash Withdrawal and Usage

- combine data from 2009, 2013 and 2017 Method of Payments Surveys
  - → survey questionnaire: demographics (age, income, location, revolver status)
  - → three-day diary: all transactions made at the POS → aggregate by day

Variable	Mean	Median	Min	Max	s.d.
Withdrawal level	141.29	100.00	0.00	1000.00	134.72
Withdrawal frequency/day	0.10	0.07	0.03	1.00	0.08
Cash holdings	63.35	40.00	0.00	500.00	70.00
Cash holdings at withdrawal	17.90	10.00	0.00	200.00	22.92
Daily total spending	84.04	50.47	0.00	977.00	105.01
Daily cash use	14.74	2.00	0.00	760.00	35.99
Cash/Total spending	0.31	0.04	0.00	1.00	0.41

Notes: Based on 3424 individual consumers across three waves (2009, 2013, 2017).

- exogenous expenditures: categorize consumers according to 30 spending types $(E[s_i], \sigma^2_{s_i})$
  - → spending type reflects total expense, *not* methods of payment used
  - → augment with consumer-specific distance measure (inverse density of branches) ▸ Details

Data: Estimation procedure I

Factual Data by Year

Variable of Interest	Overall	2009	2013	2017
Average daily cash use	14.72	15.88	15.56	12.49
Average cash withdrawal	147.80	146.25	143.54	154.99
Probability of withdrawal	0.10	0.11	0.11	0.08
Average cash holding	69.75	66.45	70.24	72.17

- model predictions given parameter vector (α, F, γ)
  - → obtained by averaging over 100×183 day simulations
  - → compare prediction vs. data to obtain error terms

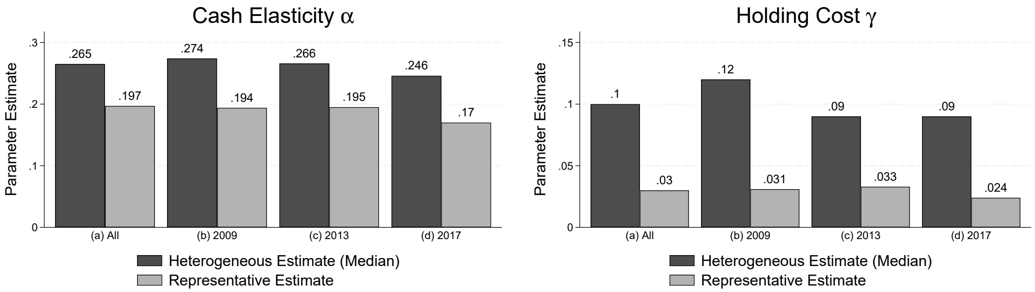
(i)	withdrawal amount given withdrawal:	$\varepsilon_{1,it} = \mathbb{E}\left[w_{it} \| w_{it} > 0\right] - \mathbb{E}\left[\widehat{w_{it} \| w_{it}} > 0\right]$
(ii)	withdrawal frequency:	$\varepsilon_{2,it} = \mathbb{E}\left[\Pr\left(w_{it} > 0\right)\right] - \mathbb{E}\left[\widehat{\Pr\left(w_{it}\right.} > 0)\right]$
(iii)	level of cash holdings:	$\varepsilon_{3,it} = \mathbb{E}\left[h_{it}\right] - \widehat{\mathbb{E}\left[h_{it}\right]}$
(iv)	level of cash use:	$\varepsilon_{4,it} = \mathbb{E}\left[c_{it}\right] - \widehat{\mathbb{E}\left[c_{it}\right]}$

  - → 12 moments via interaction with 3 instruments
    - ∗ constant term, distance measure, spending level

Data: Estimation procedure II

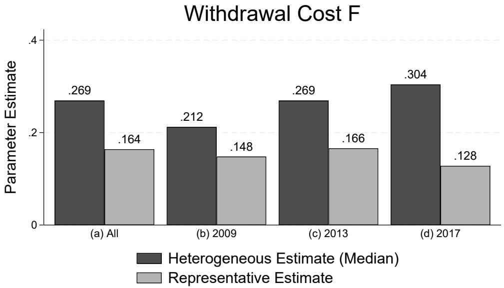
- estimation contrasts representative approach with fully heterogeneous approach
  - → representative: sample/demographic group/year-specific $(\alpha_g, F_g, \gamma_g)$
    - ∗ estimation via GMM using stacked moment conditions
    - ∗ implies optimal 2nd-stage weighting matrix
    - ! consumer spending heterogeneous (30 types), only preferences homogenous (within-group)
  - → heterogeneous: consumer-specific $(\alpha_i, F_i, \gamma_i)$
    - ∗ minimize weighted sum of squared moments (in spirit of Ackerberg 2009, Malone et al. 2021)
    - ∗ candidates: 100'000 parameter tuples obtained via Halton Draws
    - ∗ use weighting matrix from representative approach to account for different moment scales
- report results for both approaches
  - → scale holding cost γ by 100 to facilitate visualization

Takeaway I: Accounting for heterogeneity matters!



- median estimates' trend qualitatively similar to representative approach for α, γ
  - $\rightarrow$ cash elasticity and inventory cost both decline over time
  - $\rightarrow$ notable level differences

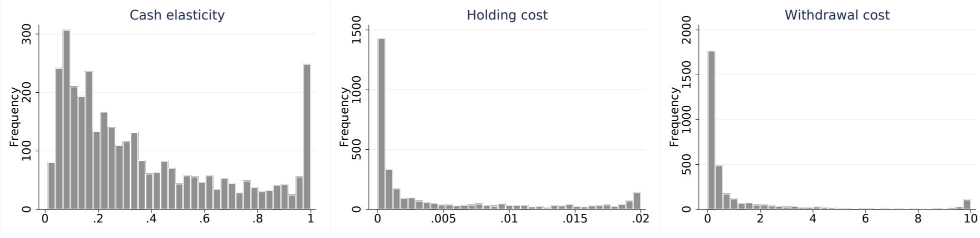
Takeaway I: Accounting for heterogeneity matters!



- qualitatively different trend for withdrawal cost parameter F
  - $\rightarrow$ upward trend for heterogeneous estimates
  - $\rightarrow$ inverse U-shaped trend for representative estimates

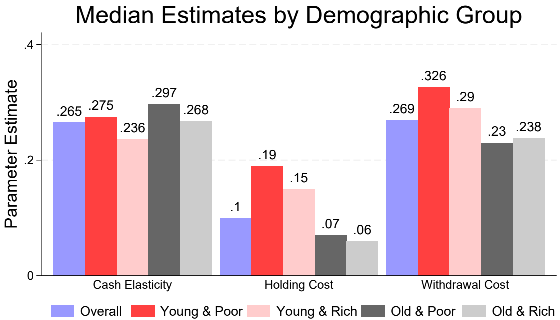
Takeaway I: Accounting for heterogeneity matters!

Distribution of α_i (left), γ_i (center), F_i (right) — All years



- heterogeneous approach reveals substantial heterogeneity
  - → cash preference α_i bi-modal but smooth between 0 and 1
  - → holding cost γ_i & withdrawal cost F_i exhibit very long tail
- model fit: heterogeneous approach outperforms representative approach
  - → true even after segmenting into finer groups (demographic group per year)
  - → representative approach unable to match increased cash holdings observed in the data

Takeaway II: Heterogeneity is partially explained by demographics!



- richer and older people use more cash but only age relation driven by preference
- older households have lower costs (both withdrawal and holding/inventory)
  - → cash elasticity increases over time for older & poorer consumers

Details

Takeaway III: Heterogeneity matters for response to Infrastructure Changes!

- evaluate factual infrastructure changes
  - → # of bank branches peaked in 2013, reflected in our sample
  - → How much cash would consumers have used in other years (relative to 2009)?

▸ Details

Infrastructure	All	Urban	Rural	Y&P	Y&R	O&P	O&R	Substantial*
2009	-	-	-	-	-	-	-	-
2013	0.99%	1.25%	-0.22%	1.51%	2.19%	0.65%	0.03%	-0.61%
2017	-0.12%	0.16%	-1.47%	-0.65%	1.80%	-0.22%	-1.08%	-15.21%

- moderate average impact of infrastructure changes on cash use on average
  - → some indications of urban/rural divide
  - → broadly follows infrastructure trend across demographic groups
    - ∗ younger, affluent consumers increase cash use
  - → 10-times larger impact following substantial changes ($\Delta d > 20\%$, 78 consumers)
    - ∗ goes hand in hand with increase in cash holdings

▸ Details

Takeaway III: Heterogeneity matters for response to Infrastructure Changes!

- evaluate factual infrastructure changes
  - → # of bank branches peaked in 2013, reflected in our sample
  - → How much cash would consumers have used in other years (relative to 2009)?

Details

Infrastructure	All	Urban	Rural	Y&P	Y&R	O&P	O&R	Substantial*
2009	-	-	-	-	-	-	-	-
2013	0.99%	1.25%	-0.22%	1.51%	2.19%	0.65%	0.03%	-0.61%
2017	-0.12%	0.16%	-1.47%	-0.65%	1.80%	-0.22%	-1.08%	-15.21%

- moderate average impact of infrastructure changes on cash use on average
  - → some indications of urban/rural divide
  - → broadly follows infrastructure trend across demographic groups
    - ∗ younger, affluent consumers increase cash use
  - → 10-times larger impact following substantial changes ($\Delta d > 20\%$, 78 consumers)
    - ∗ goes hand in hand with increase in cash holdings

Details

Takeaway III: Heterogeneity matters for response to Infrastructure Changes!

- evaluate factual infrastructure changes
  - → # of bank branches peaked in 2013, reflected in our sample ▸ Details
  - → How much cash would consumers have used in other years (relative to 2009)?

Infrastructure	All	Urban	Rural	Y&P	Y&R	O&P	O&R	Substantial*
2009	-	-	-	-	-	-	-	-
2013	0.99%	1.25%	-0.22%	1.51%	2.19%	0.65%	0.03%	-0.61%
2017	-0.12%	0.16%	-1.47%	-0.65%	1.80%	-0.22%	-1.08%	-15.21%

- moderate average impact of infrastructure changes on cash use on average
  - → some indications of urban/rural divide
  - → broadly follows infrastructure trend across demographic groups
    - ∗ younger, affluent consumers increase cash use
  - → 10-times larger impact following substantial changes ($\Delta d > 20\%$, 78 consumers)
    - ∗ goes hand in hand with increase in cash holdings ▸ Details

Takeaway III: Heterogeneity matters for response to Infrastructure Changes!

- evaluate factual infrastructure changes
  - → # of bank branches peaked in 2013, reflected in our sample
  - → How much cash would consumers have used in other years (relative to 2009)?

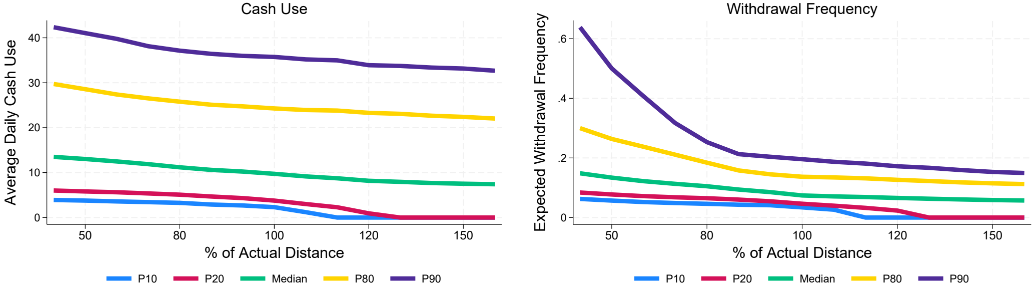
▸ Details

Infrastructure	All	Urban	Rural	Y&P	Y&R	O&P	O&R	Substantial*
2009	-	-	-	-	-	-	-	-
2013	0.99%	1.25%	-0.22%	1.51%	2.19%	0.65%	0.03%	-0.61%
2017	-0.12%	0.16%	-1.47%	-0.65%	1.80%	-0.22%	-1.08%	-15.21%

- moderate average impact of infrastructure changes on cash use on average
  - → some indications of urban/rural divide
  - → broadly follows infrastructure trend across demographic groups
    - ∗ younger, affluent consumers increase cash use
  - → 10-times larger impact following substantial changes ($\Delta d > 20\%$, 78 consumers)
    - ∗ goes hand in hand with increase in cash holdings

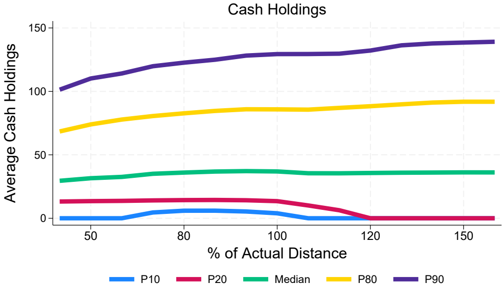
▸ Details

Takeaway III: Heterogeneity matters for response to Infrastructure Changes!



- Counterfactual: perturb distance to access-to-cash infrastructure for entire sample
  - $\rightarrow$ withdrawal costs $\Uparrow \implies$ cash use $\Downarrow$ accompanied by less frequent withdrawals
  - $\rightarrow$ some consumers stop using cash as it gets too expensive to withdraw; $\approx 10\%$ (25%) increase sufficient for 10% (20%) of consumers to stop use of cash at POS

Takeaway III: Heterogeneity matters for response to Infrastructure Changes!



- cash holdings indicative of bimodality in responses
  - → continued cash users (P80+) hold more cash to economize on withdrawals
  - → cash avoiders (P20-) stop using cash as withdrawals become too costly

Takeaway III: Heterogeneity matters for response to Infrastructure Changes!

Estimates & demographics by response to 25% increase in distance to access-to-cash infrastructure

Variable	**Cash non-users**	**Cash users**		
		all	decreased holdings	increased holdings
	(1)	(2)	(3)	(4)
Cash elasticity α	0.368	0.374	0.328	0.532
Cash holding cost γ	0.011	0.003	0.003	0.003
Withdrawal cost parameter F	2.660	1.152	0.970	1.776
Age	43.259	49.667	49.006	51.927
Income	46179.856	48063.337	47501.312	49982.079
Revolver	0.249	0.197	0.208	0.159
Urban	0.878	0.841	0.823	0.901
Young & Poor	0.342	0.251	0.265	0.203
Young & Rich	0.283	0.202	0.203	0.195
Old & Poor	0.254	0.370	0.362	0.398
Old & Rich	0.121	0.177	0.169	0.204
Observations	696	2465	1907	558

- continued cash-users (cash holdings$\Uparrow$, withdrawal amt.$\Uparrow$) are older (& more affluent)
  - $\rightarrow$ extensive margin: no difference in cash elasticity relative to cash non-users, but lower costs
  - $\rightarrow$ intensive margin: preference & costs determine response
- younger and poorer consumers phase out cash use as cash becomes too costly for them

Robustness/Extensions/Other Results

- paper: several other analyses, bigger emphasis on distributional impact
  - → branch closures largely unrelated to consumer demographics & preferences
- robustness checks
  - → different planning horizons for consumers, # of points consumer value function is solved at
  - → alternative definitions of consumer spending types and distance measures
  - → structure of the weighting matrix, application to heterogeneous approach
- conceptually feasible extensions
  - → inventory model with deposits (not on equilibrium path)
  - → estimating discounting/risk aversion (in progress)
    - ∗ heavily overidentified model (3 parameters, 12 moments)
    - ∗ evidence in favor of consumer heterogeneity (Fulford and Schuh 2017)
  - → different impact of distance, free withdrawals, . . .

Conclusion

- structural dynamic cash inventory model
  - $\rightarrow$ trade-off between more frequent withdrawals and larger cash holdings
  - $\rightarrow$ choice between cash and non-cash
  - $\rightarrow$ account for infrastructure affecting costs of withdrawal
- estimate model using detailed data on consumer behavior
  - $\rightarrow$ accounting for individual consumer heterogeneity crucial
  - $\rightarrow$ bi-modality in consumer responses to increased costs of withdrawing cash
  - $\rightarrow$ younger & less affluent consumers bear brunt of impact: cash use $\rightarrow 0$ due to cost (not preference)

Dynamic Consumer Cash Inventory Model

Kim P. Huynh
Bank of Canada

Alex Shcherbakov
Bank of Canada

André Stenzel
Bank of Canada

4th IFC Workshop on Data Science in Central Banking
Banca d'Italia
18-20 February 2025

References i

Ackerberg, D. A. (2009). A new use of importance sampling to reduce computational burden in simulation estimation. *QME*, 7:343–376.

Alvarez, F. and Lippi, F. (2009). Financial innovation and the transactions demand for cash. *Econometrica*, 77(2):363–402.

Alvarez, F. and Lippi, F. (2017). Cash burns: An inventory model with a cash-credit choice. *Journal of Monetary Economics*, 90:99–112.

Arango, C., Huynh, K. P., and Sabetti, L. (2015). Consumer payment choice: Merchant card acceptance versus pricing incentives. *Journal of Banking & Finance*, 55:130–141.

Baumol, W. J. (1952). The transactions demand for cash: An inventory theoretic approach. *The Quarterly journal of economics*, pages 545–556.

Briglevics, T. and Schuh, S. (2020). This is"what's in your wallet"... and here's how you use it. *working paper*.

Chen, H., Felt, M.-H., and Huynh, K. P. (2017). Retail payment innovations and cash usage: accounting for attrition by using refreshment samples. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 180(2):503–530.

Chen, H., Strathearn, M., and Voia, M. (2021). Consumer cash withdrawal behaviour: Branch networks and online financial innovation. Technical report, Bank of Canada.

Engert, W., Fung, B., and Segendorf, B. (2019). A tale of two countries: Cash demand in canada and sweden. Technical report, Bank of Canada.

Engert, W., Shcherbakov, O., and Stenzel, A. (2024). The impact of central bank digital currency on payments. Technical report, mimeo.

Fulford, S. and Schuh, S. D. (2017). Credit card utilization and consumption over the life cycle and business cycle. *Consumer Financial Protection Bureau Office of Research Working Paper*, (2017-03).

Greene, C. and Schuh, S. D. (2017). The 2016 diary of consumer payment choice. *Research Data Reports Paper*, (17-7).

Henry, C., Shimoda, M., and Rusu, D. (2024). 2023 methods-of-payment survey report: The resilience of cash. Technical report.

Henry, C. S., Huynh, K. P., and Welte, A. (2018). 2017 methods-of-payment survey report. Technical report, Bank of Canada Staff Discussion Paper.

Huynh, K., Molnar, J., Shcherbakov, O., and Yu, Q. (2020). Demand for payment services and consumer welfare: the introduction of a central bank digital currency. Technical report, Bank of Canada.

Huynh, K. P., Nicholls, G., and Shcherbakov, O. (2022). Equilibrium in two-sided markets for payments: Consumer awareness and the welfare cost of the interchange fee. Technical report, Bank of Canada Staff Working Paper.

Koulayev, S., Rysman, M., Schuh, S., and Stavins, J. (2016). Explaining adoption and use of payment instruments by us consumers. *The RAND Journal of Economics*, 47(2):293–325.

Lippi, F. and Secchi, A. (2009). Technological change and the households' demand for currency. *Journal of Monetary Economics*, 56(2):222–230.

References ii

Malone, J. B., Nevo, A., and Williams, J. W. (2021). The tragedy of the last mile: Economic solutions to congestion in broadband networks. *Working Paper*.

McManus, B., Nevo, A., Nolan, Z., and Williams, J. W. (2022). The steering incentives of gatekeepers in the telecommunications industry. Technical report.

Moracci, E. (2022). Cash management and payment choices revisited. *working paper*.

Nevo, A., Turner, J. L., and Williams, J. W. (2016). Usage-based pricing and demand for residential broadband. *Econometrica*, 84(2):411–443.

Rogoff, K. S. (2017). The curse of cash. In *The Curse of Cash*. Princeton University Press.

Smith, G. W. (1986). A dynamic baumol-tobin model of money demand. *The Review of Economic Studies*, 53(3):465–469.

Tobin, J. (1956). The interest-elasticity of transactions demand for cash. *The review of Economics and Statistics*, pages 241–247.

Wakamori, N. and Welte, A. (2017). Why do shoppers use cash? evidence from shopping diary data. *Journal of Money, Credit and Banking*, 49(1):115–169.

Whitesell, W. C. (1989). The demand for currency versus debitable accounts: note. *Journal of Money, Credit and Banking*, 21(2):246–251.

Model: Setup I

Return

- discrete time, $t = 0, 1, \ldots, \infty$ (infinite horizon)
- Consumers $i = 1, \ldots, N_i$:
  - $\rightarrow$ cash inventories, $h_{it-1} \geq 0$, from previous period
  - $\rightarrow$ exogenous consumption demand $s_{it} = s_i + \varepsilon_{it}$, $\varepsilon_{it} \overset{iid}{\sim} N\left(0, \sigma_{si}^2\right)$
  - $\rightarrow$ allocate payments for consumption b/w cash, $c_{it} \geq 0$, and non-cash, $s_{it} - c_{it} \geq 0$
  - $\rightarrow$ can withdraw cash, $w_{it} \geq 0$, at fixed cost depending on F_i and d_i
    - $*$ d_i: distance to branch network (data), F: scaling parameter (estimated)

$$\max_{(c_{it}, w_{it}) \forall t} \overbrace{\mathbb{E} \sum_{t=0}^{\infty} \beta^t U\left(h_{it-1}, s_{it}, d_i, w_{it}, c_{it}\right)}^{\text{discounted sum of flow utilities}}$$

$$s.t. \underbrace{0 \leq c_{it} \leq h_{it-1} + w_{it},}_{0 \,\leq\, \text{cash use} \,\leq\, \text{cash} + \text{withdrawal}} \quad \underbrace{c_{it} \leq s_{it}}_{\text{cash use} \,\leq\, \text{expenditure}} \quad , \quad \underbrace{w_{it} \geq 0}_{\text{withdrawals nonnegative}} \quad , \underbrace{h_{i,-1} = 0}_{\text{initial cash}}$$

Model: Setup II

Return

- flow utility: log-transformed Cobb-Douglas

$$u\left(h_{it-1}, s_{it}, d_i, w_{it}, c_{it}\right) = \overbrace{\alpha \ln\left(1 + c_{it}\right) + \left(1 - \alpha\right) \ln\left(1 + s_{it} - c_{it}\right)}^{\text{flow utility}} \underbrace{- F \times \mathbb{1}_{w_{it}>0} \ln(1 + d_i)}_{\text{withdrawal cost}} \underbrace{- \gamma \left(h_{it-1} + w_{it} - c_{it}\right)}_{\text{holding cost}}$$

→ α parametrizes cash preference

→ d_i reflects distance to branch network (from data)

→ F parametrizes scale of withdrawal cost

→ γ parametrizes holding cost

Model: Setup II

Return

- flow utility: log-transformed Cobb-Douglas

$$u\left(h_{it-1}, s_{it}, d_i, w_{it}, c_{it}\right) = \overbrace{\alpha \ln\left(1 + c_{it}\right) + (1-\alpha) \ln\left(1 + s_{it} - c_{it}\right)}^{\text{flow utility}} \underbrace{- F \times \mathbb{1}_{w_{it}>0} \ln(1 + d_i)}_{\text{withdrawal cost}} \underbrace{- \gamma\left(h_{it-1} + w_{it} - c_{it}\right)}_{\text{holding cost}}$$

→ α parametrizes cash preference

→ d_i reflects distance to branch network (from data)

→ F parametrizes scale of withdrawal cost

→ γ parametrizes holding cost

Model: Setup II

Return

- flow utility: log-transformed Cobb-Douglas

$$u\left(h_{it-1}, s_{it}, d_i, w_{it}, c_{it}\right) = \overbrace{\alpha \ln\left(1+c_{it}\right) + (1-\alpha)\ln\left(1+s_{it}-c_{it}\right)}^{\text{flow utility}}$$
$$\underbrace{-F \times \mathbb{1}_{w_{it}>0} \ln(1+d_i)}_{\text{withdrawal cost}} \underbrace{-\gamma\left(h_{it-1}+w_{it}-c_{it}\right)}_{\text{holding cost}}$$

→ α parametrizes cash preference

→ d_i reflects distance to branch network (from data)

→ F parametrizes scale of withdrawal cost

→ γ parametrizes holding cost

Model: Setup II

Return

- flow utility: log-transformed Cobb-Douglas

$$u\left(h_{it-1}, s_{it}, d_i, w_{it}, c_{it}\right) = \overbrace{\alpha \ln\left(1 + c_{it}\right) + (1-\alpha)\ln\left(1 + s_{it} - c_{it}\right)}^{\text{flow utility}}$$
$$\underbrace{-F \times \mathbb{1}_{w_{it}>0} \ln(1+d_i)}_{\text{withdrawal cost}} \underbrace{-\gamma\left(h_{it-1} + w_{it} - c_{it}\right)}_{\text{holding cost}}$$

  → α parametrizes cash preference

  → d_i reflects distance to branch network (from data)

  → F parametrizes scale of withdrawal cost

  → γ parametrizes holding cost

Model: Dynamics

Return

- recursive formulation of the dynamic programming problem

$$V\left(h_{it-1}, s_i, d_i, \varepsilon_{it}\right) = \max_{c_{it}\geq 0, w_{it}\geq 0} \left\{ u(h_{it-1}, s_i + \varepsilon_{it}, d_i, w_{it}, c_{it}) + \beta \int V(h_{it}, s_i, d_i, \varepsilon_{it+1}) dF_\varepsilon \right\}$$

$$s.t.\ c_{it} \leq \min\{h_{it-1} + w_{it}, s_{it}\}, h_{it} = h_{it-1} + w_{it} - c_{it}$$

→ (c_{it}, w_{it}) impact future only via (implicit) choice of cash inventory h_{it}

  * flow utility & withdrawal cost depend on (c_{it}, w_{it})
  * holding cost & future value depend on h_{it}

→ recast problem as choice of h_{it} (Δh_{it}), inducing conditionally optimal (c_{it}, w_{it})

  * analytical characterization of $(c^*_{it}(\Delta h_{it}), w^*_{it}(\Delta h_{it}))$
  * withdrawal $w_{it} > 0$ triggered by (i) depleted cash reserves or (ii) large consumption shock
  * mix of cash and non-cash throughout; withdrawal before cash is depleted

Dynamically Optimal Choice

Return

- given (w^*_{it}, c^*_{it}) we obtain $\tilde{u}(h_{it}; h_{it-1}, s_{it}) = u(h_{it-1}, s_{it}, d_i, c^*_{it}(h_{it}), w^*_{it}(h_{it}))$ and thus

$$V(h_{it-1}, s_i, d_i, \epsilon_{it}) = \max_{h_{it} \geq \max\{0, h_{it-1} - s_{it}\}} \left\{ \tilde{u}\left(h_{it}; h_{it-1}, s_{it}, d_i\right) + \beta \int V\left(h_{it}, s_i, d_i, \epsilon_{it+1}\right) dF_\epsilon \right\}$$

  $\rightarrow$ Bellman equation with current cash holdings as state & future cash holdings as control

- approximate $V(\cdot)$ by backward induction using a 183-period (6 months) planning horizon
  - $\rightarrow$ robust to perturbations in length of planning horizon, discount factor ($\beta \approx 0.95$)
  - $\rightarrow$ yields policy $h^*_{it}(h_{it-1}, s_i, d_i, \epsilon_{it})$ pinning down (w^*_{it}, c^*_{it})
  - $\rightarrow$ implied behavior in line with theoretical predictions
    - $*$ withdrawal triggered by depleted cash reserves or large expenditure s_{it}
    - $*$ costly withdrawal implies fixed post-consumption cash inventory $\tilde{h} = \arg\max_h E[V(\cdot)] - \gamma \cdot h$

Dynamically Optimal Choice

Return

- given (w_{it}^*, c_{it}^*) we obtain $\tilde{u}(h_{it}; h_{it-1}, s_{it}) = u(h_{it-1}, s_{it}, d_i, c_{it}^*(h_{it}), w_{it}^*(h_{it}))$ and thus

$$V(h_{it-1}, s_i, d_i, \epsilon_{it}) = \max_{h_{it} \geq \max\{0, h_{it-1} - s_{it}\}} \left\{ \tilde{u}\left(h_{it}; h_{it-1}, s_{it}, d_i\right) + \beta \int V\left(h_{it}, s_i, d_i, \epsilon_{it+1}\right) dF_{\epsilon} \right\}$$

  → Bellman equation with current cash holdings as state & future cash holdings as control

- approximate $V(\cdot)$ by backward induction using a 183-period (6 months) planning horizon
  → robust to perturbations in length of planning horizon, discount factor ($\beta \approx 0.95$)
  → yields policy $h_{it}^*(h_{it-1}, s_i, d_i, \epsilon_{it})$ pinning down (w_{it}^*, c_{it}^*)
  → implied behavior in line with theoretical predictions
    * withdrawal triggered by depleted cash reserves or large expenditure s_{it}
    * costly withdrawal implies fixed post-consumption cash inventory $\tilde{h} = \arg\max_h E[V(\cdot)] - \gamma \cdot h$

Model: In-period solution

Return

- ideal cash usage: $\tilde{c}(s) = \arg\max_{s \geq c \geq 0} \alpha \ln[1 + c] + (1 - \alpha) \ln[1 + s - c]$
- 3 regions depending on Δh_{it}
  - → $\Delta h_{it} < -\tilde{c}(s_{it})$: desired cash reduction exceeds ideal cash use
    - ∗ only use cash ($c_{it} = -\Delta h_{it}$), no withdrawal
    - ∗ withdrawal would be associated with even more excessive cash usage
  - → $\Delta h_{it} > 0$: increase cash holdings
    - ∗ withdrawal necessary, so $c_{it} = \tilde{c}(s_{it})$
  - → $\Delta h_{it} \in (-\tilde{c}(s_{it}), 0]$: intermediate reduction of cash holdings
    - ∗ trade-off b/w withdrawal & sub-optimal cash usage
    - ∗ either withdraw & implement ideal cash use...
    - ∗ or distort (reduce) cash use to avoid withdrawa
  - → withdrawal triggered by (i) depleted cash reserves or (ii) large consumption shock

Model: In-period solution

Return

- ideal cash usage: $\tilde{c}(s) = \arg\max_{s \geq c \geq 0} \alpha \ln[1+c] + (1-\alpha)\ln[1+s-c]$
- 3 regions depending on Δh_{it}
  - → $\Delta h_{it} < -\tilde{c}(s_{it})$: desired cash reduction exceeds ideal cash use
    - ∗ only use cash ($c_{it} = -\Delta h_{it}$), no withdrawal
    - ∗ withdrawal would be associated with even more excessive cash usage
  - → $\Delta h_{it} > 0$: increase cash holdings
    - ∗ withdrawal necessary, so $c_{it} = \tilde{c}(s_{it})$
  - → $\Delta h_{it} \in (-\tilde{c}(s_{it}), 0]$: intermediate reduction of cash holdings
    - ∗ trade-off b/w withdrawal & sub-optimal cash usage
    - ∗ either withdraw & implement ideal cash use...
    - ∗ or distort (reduce) cash use to avoid withdrawa
  - → withdrawal triggered by (i) depleted cash reserves or (ii) large consumption shock

Model: In-period solution

Return

- ideal cash usage: $\tilde{c}(s) = \arg\max_{s \geq c \geq 0} \alpha \ln[1 + c] + (1 - \alpha) \ln[1 + s - c]$
- 3 regions depending on Δh_{it}
  - → $\Delta h_{it} < -\tilde{c}(s_{it})$: desired cash reduction exceeds ideal cash use
    - ∗ only use cash ($c_{it} = -\Delta h_{it}$), no withdrawal
    - ∗ withdrawal would be associated with even more excessive cash usage
  - → $\Delta h_{it} > 0$: increase cash holdings
    - ∗ withdrawal necessary, so $c_{it} = \tilde{c}(s_{it})$
  - → $\Delta h_{it} \in (-\tilde{c}(s_{it}), 0]$: intermediate reduction of cash holdings
    - ∗ trade-off b/w withdrawal & sub-optimal cash usage
    - ∗ either withdraw & implement ideal cash use...
    - ∗ or distort (reduce) cash use to avoid withdrawa
  - → withdrawal triggered by (i) depleted cash reserves or (ii) large consumption shock

Model: In-period solution

Return

- ideal cash usage: $\tilde{c}(s) = \arg\max_{s \geq c \geq 0} \alpha \ln[1 + c] + (1 - \alpha) \ln[1 + s - c]$
- 3 regions depending on Δh_{it}
  - → $\Delta h_{it} < -\tilde{c}(s_{it})$: desired cash reduction exceeds ideal cash use
    - * only use cash ($c_{it} = -\Delta h_{it}$), no withdrawal
    - * withdrawal would be associated with even more excessive cash usage
  - → $\Delta h_{it} > 0$: increase cash holdings
    - * withdrawal necessary, so $c_{it} = \tilde{c}(s_{it})$
  - → $\Delta h_{it} \in (-\tilde{c}(s_{it}), 0]$: intermediate reduction of cash holdings
    - * trade-off b/w withdrawal & sub-optimal cash usage
    - * either withdraw & implement ideal cash use...
    - * or distort (reduce) cash use to avoid withdrawa
  - → withdrawal triggered by (i) depleted cash reserves or (ii) large consumption shock

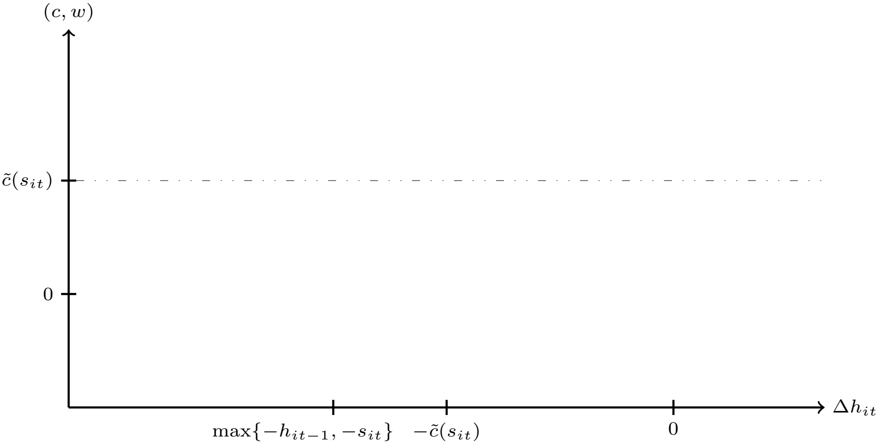
Model: In-period solution

Return

- ideal cash usage: $\tilde{c}(s) = \arg\max_{s \geq c \geq 0} \alpha \ln[1+c] + (1-\alpha)\ln[1+s-c]$
- 3 regions depending on Δh_{it}
  - → $\Delta h_{it} < -\tilde{c}(s_{it})$: desired cash reduction exceeds ideal cash use
    - ∗ only use cash ($c_{it} = -\Delta h_{it}$), no withdrawal
    - ∗ withdrawal would be associated with even more excessive cash usage
  - → $\Delta h_{it} > 0$: increase cash holdings
    - ∗ withdrawal necessary, so $c_{it} = \tilde{c}(s_{it})$
  - → $\Delta h_{it} \in (-\tilde{c}(s_{it}), 0]$: intermediate reduction of cash holdings
    - ∗ trade-off b/w withdrawal & sub-optimal cash usage
    - ∗ either withdraw & implement ideal cash use...
    - ∗ or distort (reduce) cash use to avoid withdrawa
  - → withdrawal triggered by (i) depleted cash reserves or (ii) large consumption shock

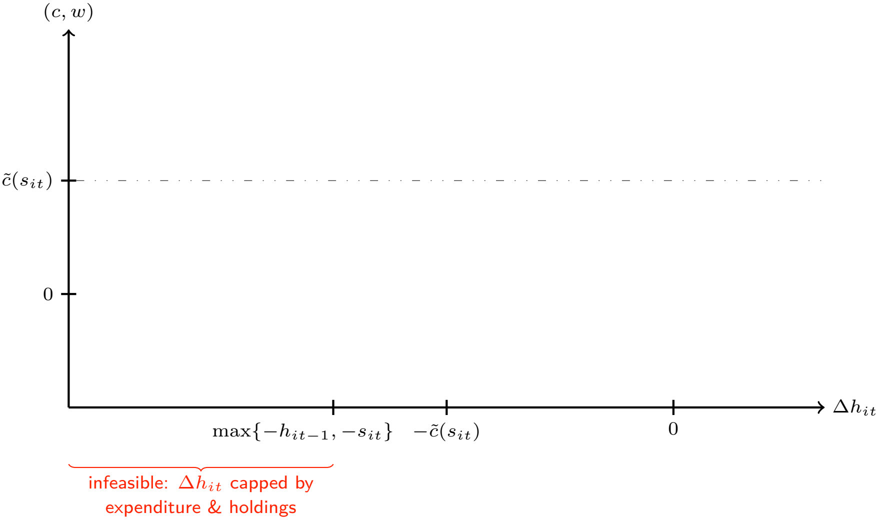
Model: In-period solution

◂ Return



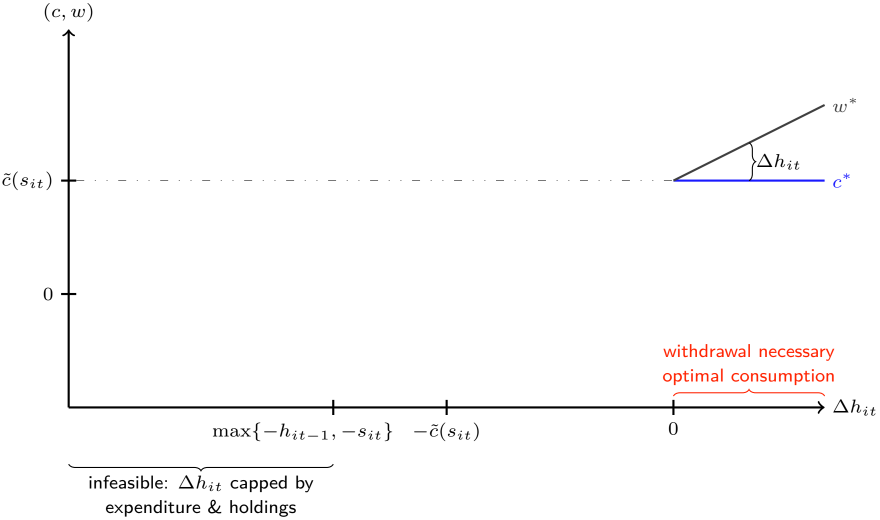
Model: In-period solution

Return



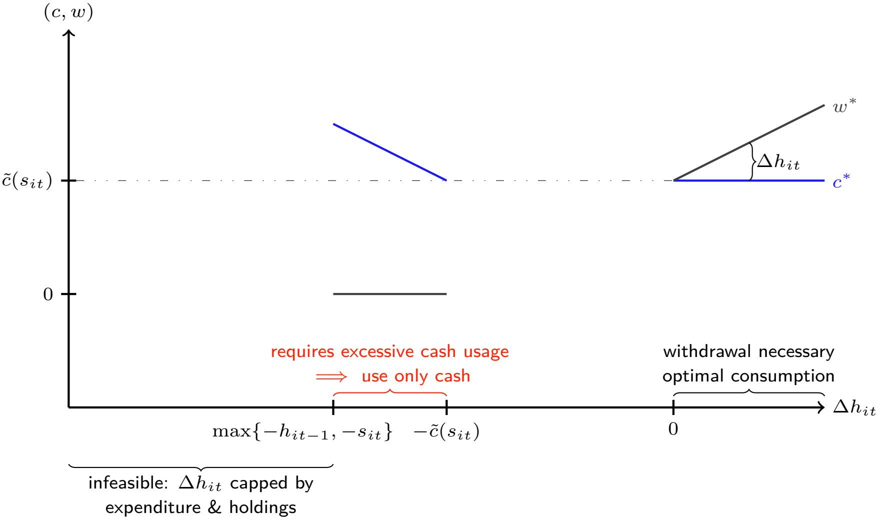
Model: In-period solution

Return



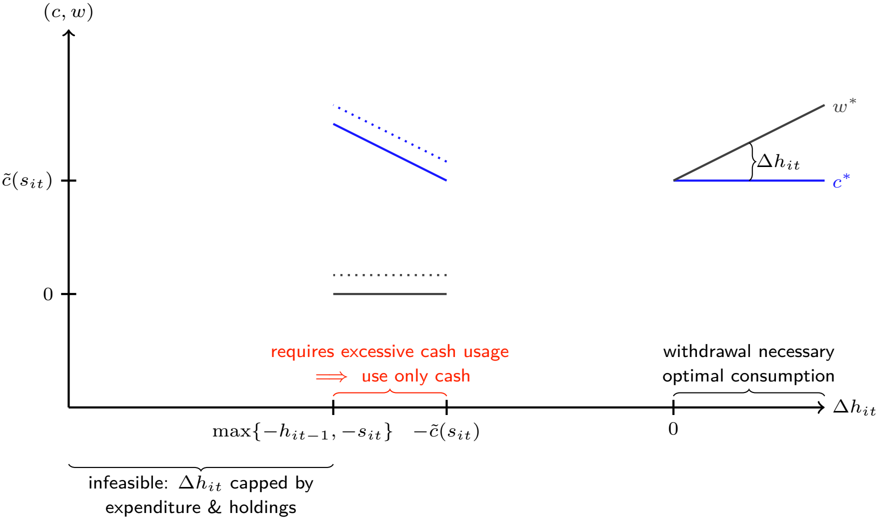
Model: In-period solution

Return



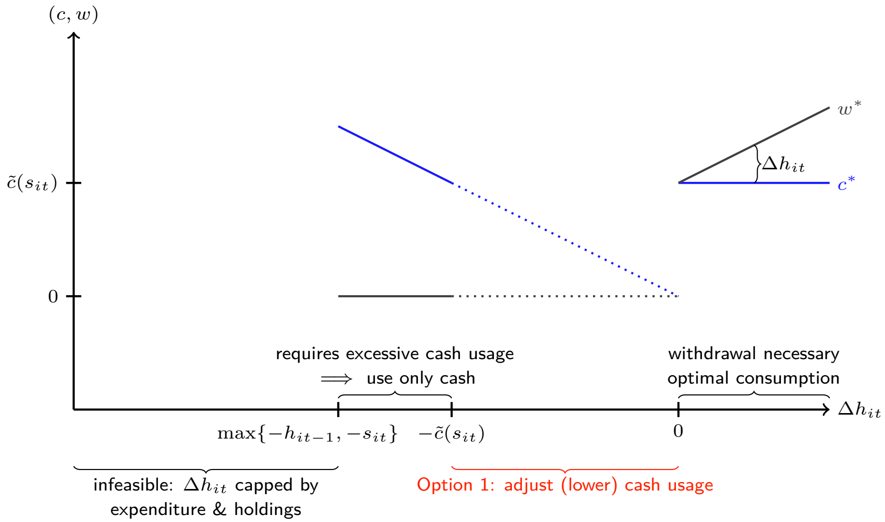
Model: In-period solution

Return

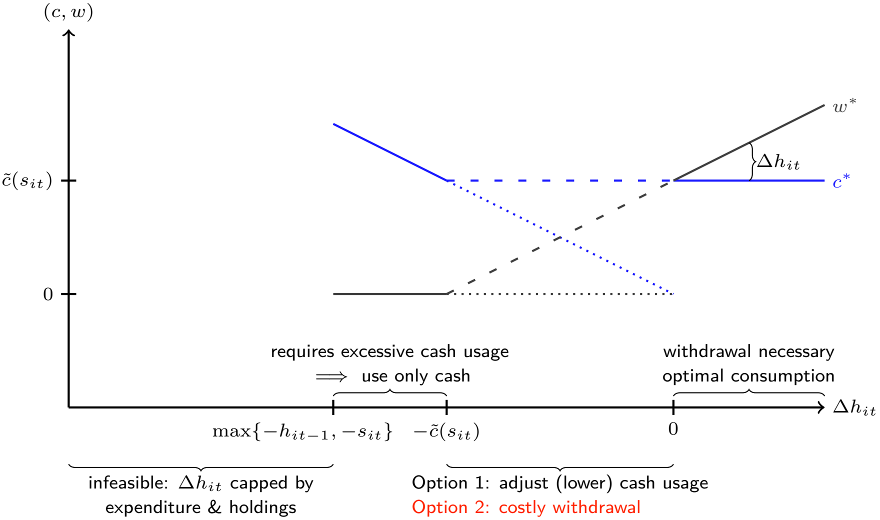


Model: In-period solution

Return

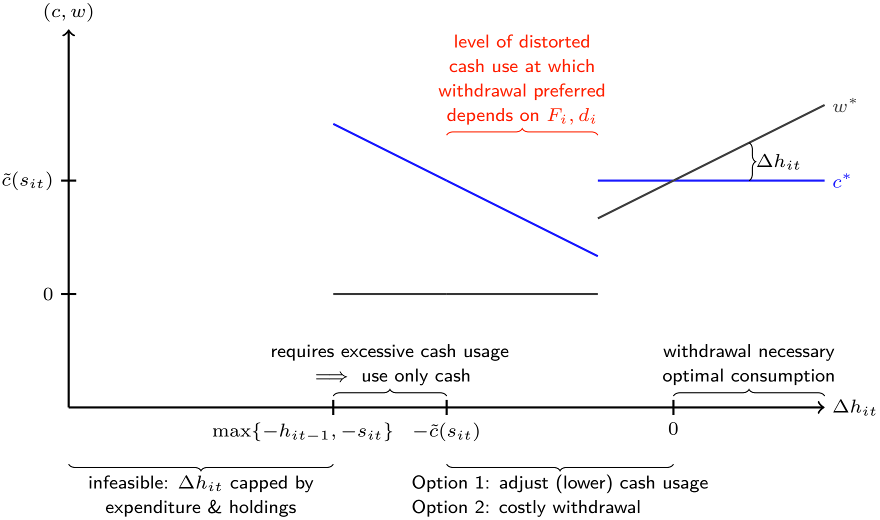


Model: In-period solution



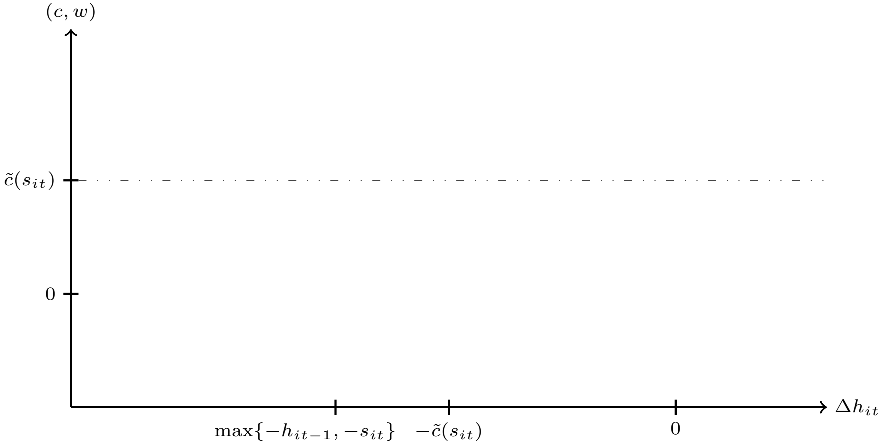
Model: In-period solution

Return



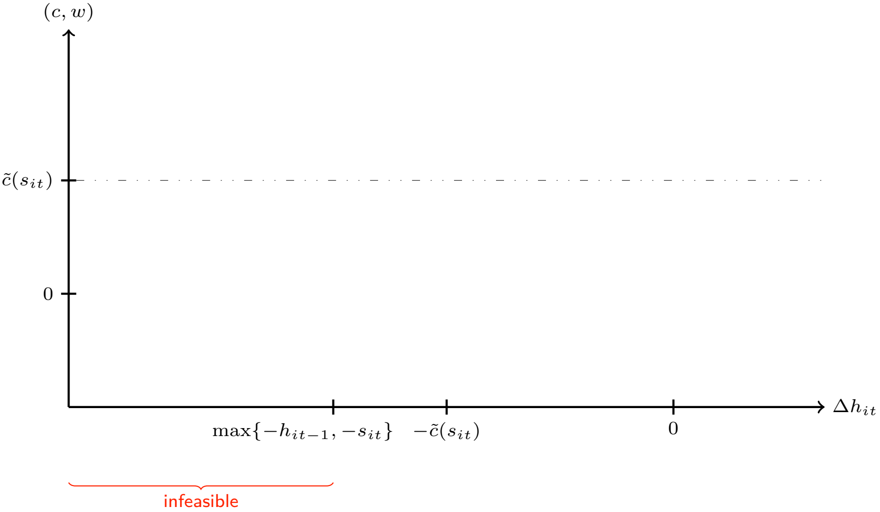
Model: In-period solution:

Return



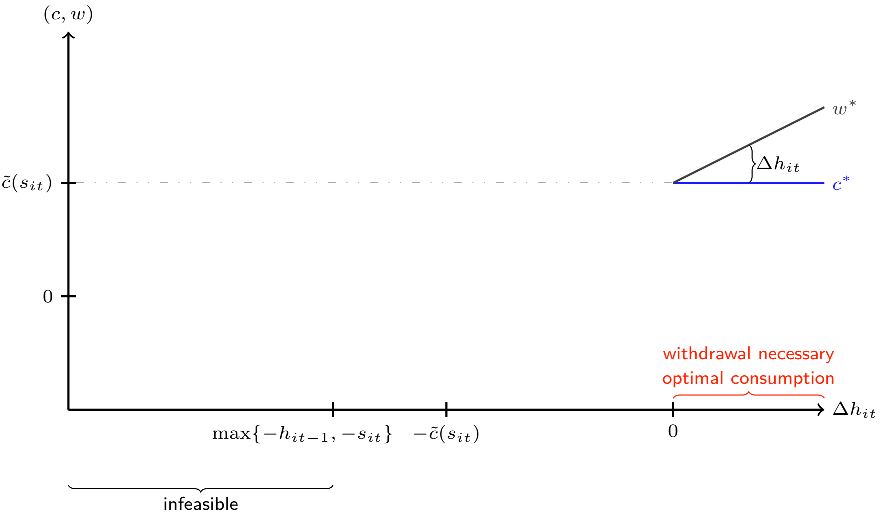
Model: In-period solution:

Return



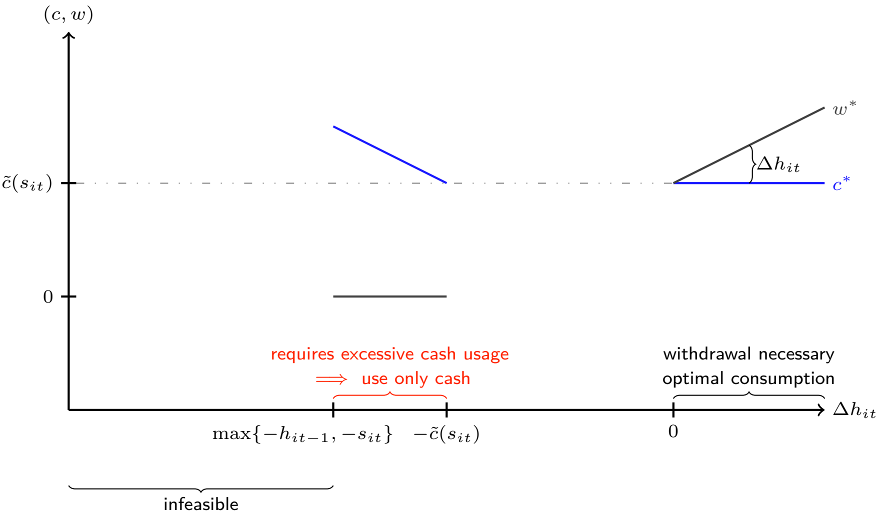
Model: In-period solution:

Return



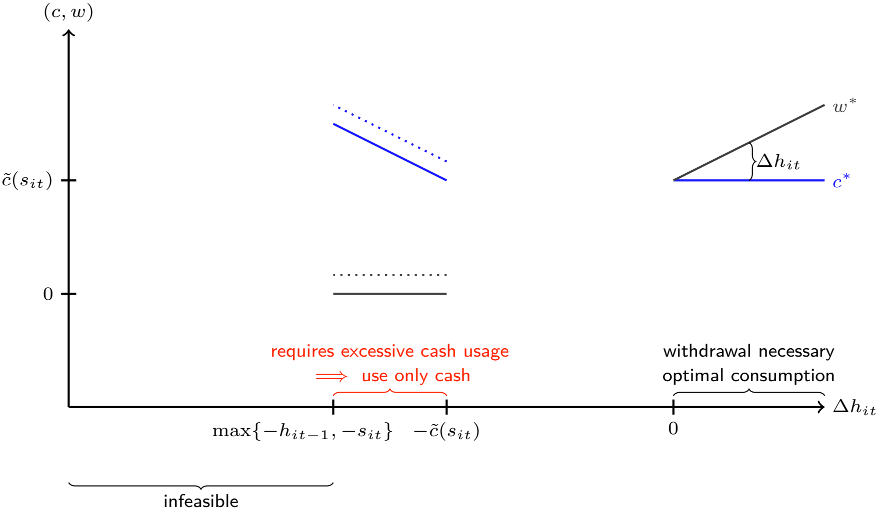
Model: In-period solution:

Return



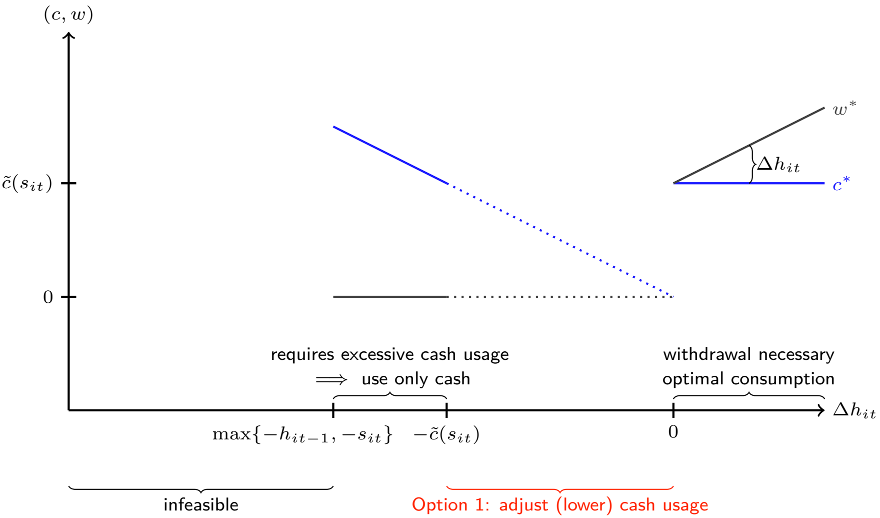
Model: In-period solution:

Return



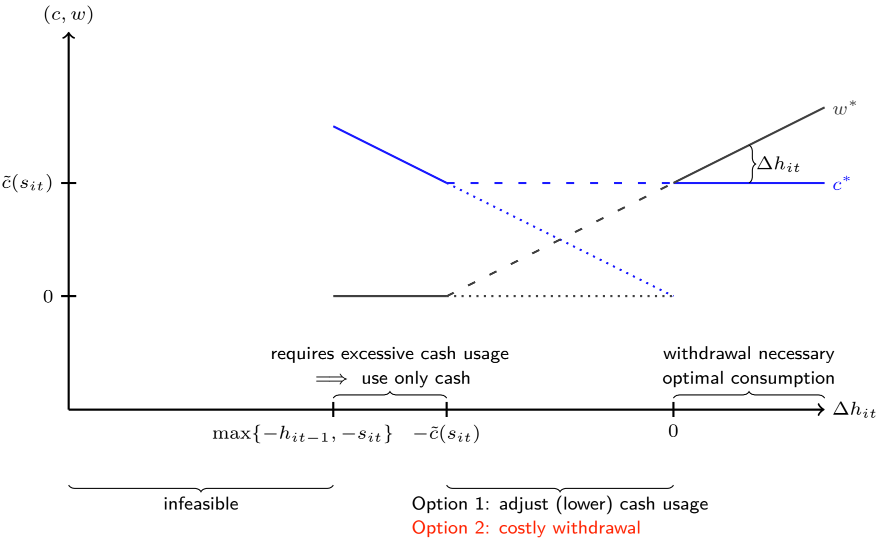
Model: In-period solution:

Return



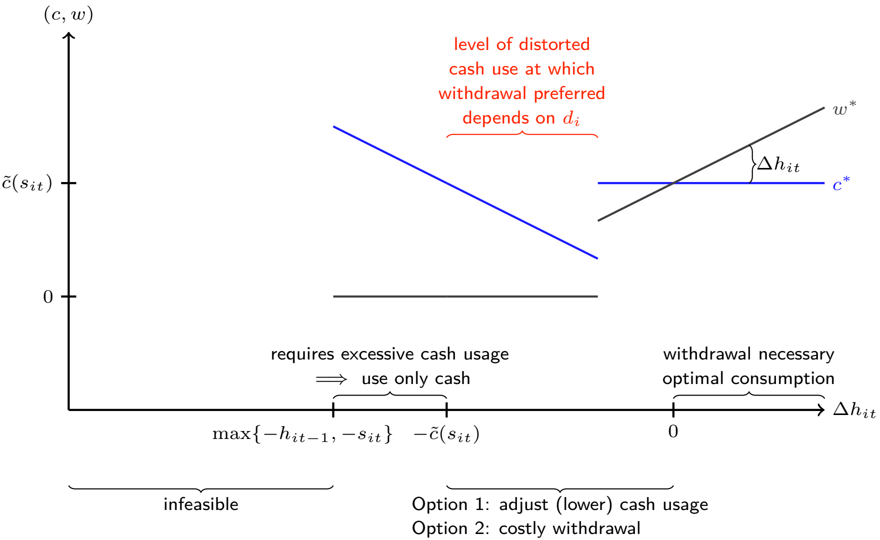
Model: In-period solution:

Return



Model: In-period solution:

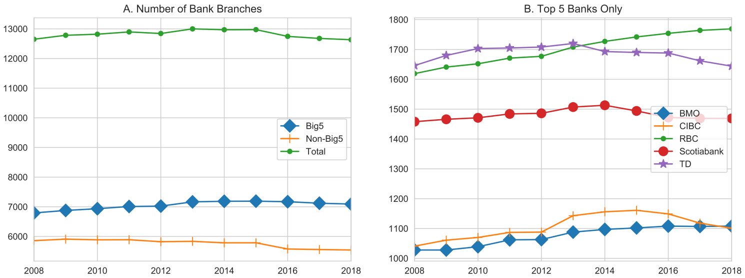
Return



Data: Distance to Bank Network

Return

- d_i reflects density of bank branches in forward sortation area (FSA, 1643 in CAN)
  - $\rightarrow$ $d_i = \frac{\ln(\text{geographic area of FSA})}{\#\text{ of banch branches}}$
  - $\rightarrow$ ATMs by FIs typically co-located with branches (88%)
  - $\rightarrow$ white-label ATMs only account for $\approx$ 20% of ATM withdrawals (Chen et al. 2021)



Evolution of bank branches.

Evolution over time by Demographic Group

Return

Table 1: Evolution of parameter estimates by demographic type (median estimates)

	Young & Poor	Young & Rich	Old & Poor	Old & Rich	Urban	Rural
(a) Cash elasticity, α						
2009	0.282	0.240	0.269	0.311	0.259	0.344
2013	0.282	0.216	0.297	0.232	0.257	0.333
2017	0.245	0.235	0.297	0.234	0.245	0.257
(b) Holding cost, γ						
2009	0.00161	0.00162	0.00062	0.00091	0.00112	0.00153
2013	0.00209	0.00150	0.00072	0.00054	0.00080	0.00255
2017	0.00154	0.00134	0.00077	0.00049	0.00082	0.00145
(c) Withdrawal cost, F						
2009	0.257	0.258	0.141	0.209	0.269	0.084
2013	0.298	0.276	0.240	0.223	0.290	0.105
2017	0.394	0.392	0.269	0.296	0.375	0.100
N-obs.	924	749	1189	562	2902	522

Evolution of Infrastructure (Sample)

Return

Year	Mean	p10	p25	Median	p75	p90	N-obs.
2009	288.14	0.71	1.67	6.35	111.66	573.90	973
2013	160.32	0.52	1.24	3.42	40.72	279.70	1408
2017	205.65	0.48	1.24	3.48	38.72	287.62	1043
All	210.50	0.57	1.38	3.95	55.62	349.48	3424

Notes: The distance measure is the natural logarithm of the geographic area of the forward sortation area (FSA) consumers are located in in square kilometers over the number of available bank branches.

Evaluation of Factual Infrastructure Changes: Cash Holdings

Return

Infrastructure	All	Urban	Rural	Y&P	Y&R	O&P	O&R	Substantial*
2009	-	-	-	-	-	-	-	-
2013	0.18%	0.23%	-0.13%	1.16%	0.16%	-0.18%	-0.06%	0.75%
2017	0.13%	0.23%	-0.35%	0.45%	0.70%	-0.15%	-0.09%	2.69%

Elasticity

◂ Return

Elasticity of predictions w.r.t. increase in distance (25% change)

(a) All consumers	mean	min	p10	p25	p50	p75	p90	max
average withdrawal amount	-0.75	-4.00	-4.00	-0.01	0.01	0.09	0.49	2.32
expected withdrawal frequency	-1.19	-4.00	-4.00	-2.06	-0.34	-0.12	-0.05	0.00
average cash holding	-0.61	-4.00	-4.00	-0.13	-0.02	0.00	0.37	3.59
average cash use	-1.09	-4.00	-4.00	-1.81	-0.19	-0.06	-0.01	0.00
expected payoff per period	-0.14	-3.98	-0.30	-0.08	-0.01	-0.00	-0.00	-0.00
Observations	3161							

(b) Cash users (post increase)	mean	min	p10	p25	p50	p75	p90	max
average withdrawal amount	0.16	-0.28	0.00	0.00	0.01	0.24	0.56	2.32
expected withdrawal frequency	-0.39	-3.58	-0.73	-0.44	-0.23	-0.08	-0.04	0.00
average cash holding	-0.00	-3.48	-0.17	-0.06	-0.02	0.00	0.46	3.59
average cash use	-0.27	-3.48	-0.48	-0.27	-0.12	-0.04	-0.01	0.00
expected payoff per period	-0.05	-2.65	-0.19	-0.03	-0.01	-0.00	-0.00	-0.00
Observations	2465							

(c) Cash non-users (post increase)								
expected payoff per period	-0.46	-3.98	-1.46	-0.37	-0.07	-0.00	-0.00	-0.00
Observations	696							

- focus on 25% change
  - → reaffirm bi-modality
  - → 22% no longer use cash
- Who does what?