CLOUD-CG: Clustering on Longitudinal Causal Graphs

Shinpei Nakamura-Sakai[1, 2*], **Yuhe Gao**[3 *] **Chi-Hui Yen**[3],
Christoph Scheidiger[3†], **Jasjeet Sekhon**[1 †]

[1] Yale University, [2] Banco de México, [3] Amazon
{s.nakamura.sakai, jasjeet.sekhon}@yale.edu
{gaoyuhe, chihuyen, scheic}@amazon.com

Abstract

To the best of our knowledge, this work introduces the first framework for clustering longitudinal data by leveraging time-dependent causal representation learning. Clustering longitudinal data has gained significant attention across various fields, yet traditional methods often overlook the causal structures underlying observed patterns. Understanding how covariates influence outcomes is critical for policymakers and business leaders seeking actionable and interpretable insights. Although causal discovery models have advanced from static to time-series frameworks, their integration with longitudinal data remains underexplored. To address this limitation, we propose CLOUD-CG (Clustering on Longitudinal Causal Graphs), a method that leverages Temporal Directed Acyclic Graphs (T-DAGs), to cluster longitudinal data based on causal mechanisms represented by T-DAGs. CLOUD-CG preserves unit-level heterogeneity, enabling the identification of groups with similar causal structures and delivering interpretable insights. We validate the framework through extensive simulations and demonstrate its practical utility by applying it to deposit data from commercial banks in Mexico. This application reveals how macroeconomic variables causally influence deposits, providing policymakers with a robust tool to monitor and enhance financial stability in emerging markets. Our work contributes to the growing field of clustering and causal discovery within longitudinal data analysis, offering new possibilities for understanding complex, time-dependent relationships across various domains.

Introduction

Clustering longitudinal data has garnered significant attention in recent years, with applications spanning diverse fields such as social science (Huang et al. 2011), business (Ha, Bae, and Park 2002), and environmental studies (Beckers et al. 2020). Notably, it has been widely adopted in medicine and healthcare for phenotyping and diagnosing disease progression patterns (Ali et al. 2021; Birkenbihl et al. 2023). However, most existing methods focus narrowly on specific outcome variables and fail to consider the broader, interconnected structure of covariates. For example, two recent review papers on longitudinal data clustering (Lu 2024; Den Teuling, Pauws, and van den Heuvel 2021) discuss a range of models but largely emphasize the trajectories of individual outcome variables, neglecting scenarios involving complex relationships between multiple covariates. In many real-world applications, the goal is not merely to analyze trajectories but to identify cohorts influenced by multiple treatments, mediated through a defined set of intermediate variables and resulting in multiple outcome variables. Capturing these intricate causal relationships is essential for advancing insights in longitudinal data analysis.

In the past decade, causal discovery models have gained significant traction in identifying key drivers across various domains, from economics to epidemiology. The evolution of these models has seen a shift from static approaches, such as the Peter-Clark (PC) algorithm (Spirtes, Glymour, and Scheines 2000) and NOTEARS (Zheng et al. 2018), to more sophisticated time-series discovery methods, exemplified by the PCMCI+ algorithm (Runge 2020). PCMCI+ algorithm extends the PC algorithm to time-series settings by integrating the Momentary Conditional Independence (MCI) test. However, when the data available are richer, containing trajectories of time series data from different units, there is no direct way to apply time-series causal discovery algorithms. One may aggregate such longitudinal data over the dimension of different units to convert it into a single causal trajectory and apply PCMCI+ or other algorithms to it. However, such a procedure causes significant information loss on heterogeneous causal relationships among units.

One way to address this gap is by learning causal graphs for each individual unit in a longitudinal dataset. However, in real-world scenarios, the number of trajectories in a longitudinal dataset can be immense, making presenting all the individual causal graphs to decision-makers impractical. Such an approach could potentially obscure the key data signals. It is worth noting that longitudinal data with numerous trajectories often exhibit clustering structures. This observation motivates us to perform clustering on temporal causal graphs learned from individual trajectories.

To address these challenges, we introduce the CLOUD-CG (Clustering on Longitudinal Causal Graphs), designed to learn temporal causal graphs from longitudinal data and recover the latent cluster structures within heterogeneous causal relationships. Specifically, our approach leverages Temporal Directed Acyclic Graphs (T-DAGs) to represent

*Denotes co-first authorship.
†Denotes co-senior authorship.

the causal relationships between variables. We propose the first clustering algorithm for T-DAGs to group individuals with similar causal dynamics into clusters, by leveraging a novel distance metric between T-DAGs, the Temporal Structural Hamming Distance (T-SHD).

T-SHD extends the traditional Structural Hamming Distance (SHD) used in static causal discovery by capturing both the adjacency information in causal graph topology and the time lags of causal effects, which are unique to temporal causal discovery. Unlike SHD, T-SHD introduces differential penalties for errors in edge time lags and covariate adjacency mismatches during the clustering process. Hence, this approach enables more precise predictions of cluster numbers and labels, enhancing the accuracy of clustering outcomes.

We demonstrate the credibility of our method by conducting two simulation studies. In the first simulation study, we focus on the clustering algorithm's ability to recover true cluster structures. We propose a novel simulation procedure to randomly generate synthetic T-DAGs with various settings of cluster numbers, node numbers, time lags, as well as distances between/within clusters. Our results demonstrate that the proposed clustering method consistently achieve high clustering performance across various configurations, as measured by the Adjusted Rand Index (ARI) against the true underlying cluster structure. In the second simulation study, we evaluate the performance of the entire CLOUD-CG framework, where clustering is performed with the presence of T-DAG estimation noise. In this study, rather than clustering on simulated T-DAGs, we first apply a causal discovery algorithm to estimate the temporal causal graph from simulated data, then cluster on those estimated T-DAGs. The results indicate that our clustering method maintains high accuracy, even in the presence of noise introduced during the learning procedure of T-DAGs. These extensive simulation studies provide strong evidence of the robustness of our proposed clustering algorithm across various cluster and graphical structures, as well as under different levels of graph estimation noise.

To demonstrate the real-world applicability of our method, we apply it to a comprehensive longitudinal dataset at the bank level within the Mexican banking market. This dataset spans 19 years (2006–2024) and includes monthly observations for 49 banks. It features data on bank deposits alongside key variables commonly used in macroeconomic analysis, such as the Mexican Stock Market Index, exchange rates, interest rates, inflation, unemployment, and economic indicators for the United States. We use this dataset to investigate the key factors influencing bank deposits and to uncover how different banks respond to these macroeconomic factors. Our analysis identifies distinct clusters of banks characterized by varying driving factors and response patterns. Our results show that smaller banks are vulnerable to market volatility, larger institutions are sensitive to unemployment and equity trends, and non-national banks show resilience to shocks.

Our contributions

- We extend Time Series Causal Discovery to longitudinal data settings, uncovering heterogeneous causal relationships across units.
- We propose the first clustering method for longitudinal data leveraging time-dependent causal representation learning, demonstrating robustness to diverse graphical structures, cluster configurations, and causal discovery errors through extensive simulation experiments.
- We provide insights into the factors driving deposits at the institutional level using data from Mexican commercial banks, highlighting heterogeneities that are often hidden in aggregate analyses.

Clustering on Longitudinal Causal Graphs: CLOUD-CG

Preliminaries

We focus on the PCMCI+ method (Runge 2020) among various TSCD methods as our TSCD base algorithm, as it allows for contemporaneous edges and provides a T-DAG that quantifies the specific time lags at which variables impact the outcome. This capability is crucial for monitoring financial stability, enabling policymakers to take timely actions to preserve the financial system's health. PCMCI+ makes several important assumptions: Causal Sufficiency, Causal Markov Condition, Adjacency Faithfulness Conditions, Consistent Conditional Independence Test, and Causal Stationarity. Most of these assumptions are commonly imposed in static causal discovery (Spirtes, Glymour, and Scheines 2000).

In particular, causal stationarity refers to the assumption that the graph structure remains consistent across all time points. Formally,

Assumption 1 (Causal Stationarity) *For two variables $X^i_{t-\tau}$ and X^j_t in the temporal data trajectory, if causal relation $X^i_{t-\tau} \to X^j_t$ holds at some time pointt, then $X^i_{t'-\tau} \to X^j_{t'}$ for all $t' \neq t$ where $t, t' \in \{1, 2, \cdots, T\}$,*

where superscripts i, j denote covariate indices, while the subscript t represents the time point, X refers to an entry within the discrete-time structural causal system of M variables $\mathbf{X}_t = (X^1_t, ..., X^M_t)$, and T is the entire time horizon of X.

This assumption is especially important in our work, as it enables the creation of T-DAGs for each unit within the longitudinal data, with a specified maximum time lag, τ_{max}. The resulting T-DAGs for each unit form a tensor with dimensions $(M, M, \tau_{max} + 1)$.

Multiplex T-DAG

We leverage repeated measurements of units and Assumption 1 to obtain a T-DAG for each unit in the longitudinal dataset to generate $\{G_i\}_{i=1}^N$, where each G_i is the T-DAG for unit i, collectively forming a multiplex T-DAG. Figure 1 illustrates the distribution concept for these T-DAGs, providing a visual representation of this framework.

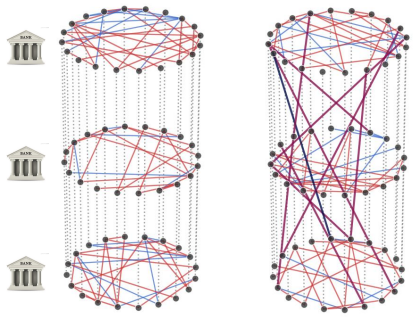


Figure 1: Visualizing the Multiplex T-DAGs under two scenarios: with independent inter-unit relationships (left) and with inter-unit dependencies (right). Each layer represents a T-DAG for a unit, where nodes are covariates and edges reflect the learned causal structure. Vertical dotted lines indicate the nodes across T-DAGs corresponding to the same variables.

TSCD itself cannot be directly applied to the longitudinal data consisting of multiple units because it is designed for single time-series trajectories rather than datasets with multiple units. To address this limitation, the longitudinal data can be aggregated by taking the mean and quantiles across all units to transform them into time series data of a single trajectory. However, such aggregation reduces the data size by a factor of $1/N$, and interpreting the aggregated data can be non-intuitive, such as understanding a variable's effect on the outcome's quantile. By studying the full distribution of T-DAGs, we preserve the heterogeneity of the units, allowing us to extract more relevant insights. Moreover, in complex causal systems where causal sufficiency assumptions may not hold, including unit-level information can help recover unobserved confounders that may not be explicitly visible in the data.

Temporal Structural Hamming Distance: T-SHD

To cluster units in longitudinal data using the estimated multiplex-T-DAG, we need to define a distance metric suitable for T-DAGs. Traditional Structural Hamming Distance (SHD) is designed for static DAGs without temporal information, so directly adopting this metric may not improve clustering performance in time-series context. Therefore, we introduce a new distance metric, the Temporal Structural Hamming Distance (T-SHD), specifically designed for T-DAGs.

Definition 1 (Temporal-Structural Hamming Distance) *The Temporal-Structural Hamming Distance (T-SHD) between T-DAGs G_1 and G_2 of dimention $\mathbb{N}^{\mathcal{M}\times\mathcal{M}\times\mathcal{T}}$ is defined as:*

$$TSHD = \sum_{(i,j,t)\in\mathbb{N}^{\mathcal{M}\times\mathcal{M}\times\mathcal{T}}} w_{(i,j,t)}|v_1(i,j,t) - v_2(i,j,t)|$$

Here, the inputs to T-SHD are generalized to weighted adjacency matrices of two graphs, where $v_g(i,j,t)$ represents the edge weight from source node i to target node j with time-lag t in graph $g \in \{1,2\}$. The edge weights can be obtained by calculating the partial correlation between the two nodes, fitting the functional regression model of parent-children nodes, or learning the a conditional probability distribution. When $v_g(i,j,t)$ is binary, it reduces to a traditional adjacency matrix. The weighting hyper-parameter $w_{(i,j,t)}$ adjusts the contribution of each edge based on its presence and time lag, which is defined as:

$$w_{(i,j,t)} = \begin{cases} \tau \cdot w_t, & \text{if } (i,j) \in A, \\ w_n, & \text{if } (i,j) \notin A, \end{cases}$$

where:

$$A = \{(i,j) \mid v_1(i,j,t_1) \neq 0 \text{ and } v_2(i,j,t_2) \neq 0 \text{ for some } t_1, t_2\}.$$

The parameter τ is the minimum time lag difference between t_1 and t_2 for edges in A. Here, A is the set of node tuples that have directed edges in both graphs G_1 and G_2 , irrespective of the time lag. The terms w_t and w_n are hyperparameters that control the weights assigned to time-lag mismatches and covariate mismatches respectively. When $w_t = 0$, the T-SHD reduces to the standard SHD by ignoring the temporal dimension mismatch. This highlights that T-SHD is a generalization of SHD, providing the flexibility to penalize errors in both time-lag differences and adjacency mismatches between T-DAGS. In general, we recommend ensuring that $\tau_{\max} \cdot w_t < w_n$, as mistakes in edge presence or absence are typically more consequential than discrepancies in the number of lags. Additional tuning details of these two hyperparameters are in the next section.

CLOUD-CG

In this section, we introduce CLOUD-CG, a method designed to cluster longitudinal causal graphs. The input for this method consists of multiplex T-DAGs, and the primary goal is to group units whose causal structures, as represented in the T-DAGs, are similar.

First, we perform multiple T-DAG estimation using PCMCI+ on the longitudinal data $\mathcal{D}$. As discussed earlier, this estimation can be carried out under two settings: one without inter-unit relationships and one with inter-unit relationships. The choice of setting depends on the structure of the data and prior knowledge. For instance, if the data is related to bank transfers between entities, the setting with inter-unit relationships should be chosen, as there are interactions between units.

While incorporating inter-unit relationships increases the complexity of the algorithm, this cost can be mitigated by leveraging prior knowledge about the interactions. Such prior knowledge can be specified using a tensor $\mathcal{P}$ with dimensions $(\mathcal{M}, \mathcal{M}, \mathcal{T})$, where each entry indicates either the presence of an edge (including its direction) or the absence of an edge. In addition, two critical hyperparameters must be considered: the maximum time lag, $\tau_{\max}$, and the significance threshold, α. Detailed guidance on selecting these parameters is provided in Runge (2020).

The next step is T-DAG filtering. When clustering T-DAGs, researchers may be interested in focusing on specific aspects of the causal system. For instance, they might analyze the effects on a particular set of end nodes, denoted as $\mathcal{E}$, the impact caused by a specific set of starting nodes, denoted as $\mathcal{S}$, or the relationships between both. To achieve this, we employ a breadth-first search algorithm, which allows us to isolate and explore the relevant parts of the T-DAG. By narrowing the focus to these specific regions of interest, we can gain deeper insights into targeted sections of the causal system, rather than clustering based on the entire T-DAG.

Optionally, we can tune the hyperparameters—the number of clusters, k, and the realization weight on time lag, w_t, by selecting values from predefined sets K (possible cluster numbers) and W_t (time lag weights), if these values are not specified beforehand. To perform the tuning, we calculate an unsupervised loss function, $\mathcal{L}_{\text{clust}}$, which evaluates the quality of clustering. Common choices for $\mathcal{L}_{\text{clust}}$ include the Silhouette score (Rousseeuw 1987) and the elbow method (Thorndike 1953).

Once we get the multiplex T-DAGs estimated, hyperparameters defined, and relevant nodes filtered, we calculate the T-SHD for every pair of the N units to form a two-dimensional distance matrix $\mathbf{Z} \in \mathbb{R}^{N \times N}$. We can then apply a clustering method $f_{\text{clust}} : (\mathbf{Z}, k) \rightarrow \mathbf{C}$ to perform the clustering on the T-SHD matrix $\mathbf{Z}$, and obtain the cluster label vector $\mathbf{C} = \{c_1, .., c_N\}$ with $c_i \in \{1, ..., k\}$. The clustering method f_{clust} can be any clustering algorithm that accepts a pair-wise distance matrix as input, such as hierarchical clustering, DBSCAN, or similar techniques. The full details of the algorithm is in Algorithm 1.

Algorithm 1: CLOUD-CG

Input: $\mathcal{D}$, f_{clust}, $\mathcal{L}_{\text{clust}}$, τ_{max}, α, $\mathcal{S}(optional)$, $\mathcal{E}(optional)$, $\mathcal{P}(optional)$, $k(optional)$, $w_t(optional)$
Estimate $\{G\}_{i=1}^{N} = PCMCI^{+}(\mathcal{D}, \tau_{max}, \alpha, \mathcal{P})$ with or without inter-unit relationship
if k or w_t are not predefined **then**
    for $(k, w_t) \in K \times W_t$ **do**
        Create $\{G_i'\}_{i=1}^{N}$ by T-DAG filtering using $\mathcal{S}, \mathcal{E}$
        Compute $Z_{i,j}^{(k,w_t)} = T\text{-}SHD(G_i, G_j, w_t)$ for all (i, j) to construct $\mathbf{Z}^{(k,w_t)}$
        Compute $\mathbf{C}^{(k,w_t)} = f_{\text{clust}}(\mathbf{Z}^{(k,w_t)}, k)$
        Compute $\mathcal{L}_{\text{clust}}^{(k,w_t)}(\mathbf{C}^{(k,w_t)}, \mathbf{Z}^{(k,w_t)})$
    end for
end if
return $\mathbf{C}^{(k^*, w_t^*)}$ where

$$(k^*, w_t^*) = argmin_{(k,w_t)} \mathcal{L}_{\text{clust}}^{(k,w_t)}$$

Simulation Study

We perform two simulation studies to investigate the performance of the proposed clustering algorithm. In the first simulation, we investigate whether the CLOUD-CG algorithm can recover the true cluster labels on randomly generated T-DAGs. Our results show the clustering algorithm is robust to various graph parameters and cluster structures. In the second simulation study, we investigate the performance of the proposed CLOUD-CG algorithm on estimated T-DAGs, where the true underlying T-DAGs are unknown. That is, the causal graphs are first estimated by performing time series causal discovery on longitudinal data for each unit. Then, we run the clustering algorithm to assign a cluster label to each estimated T-DAG. The results demonstrate that the proposed framework returns accurate clustering results in most settings despite the estimation errors in causal discovery.

Simulation study 1: Performance of Clustering on true T-DAGs

We first specify the generating procedure of simulated clusters of T-DAGs. In the simulation, cluster centers are randomly created, then the T-DAGs belonging to each cluster are generated around the cluster centers. When generating k clusters, each cluster center is a T-DAG with M nodes following the Erdős–Rényi (ER) model. Specifically, edges between nodes are independently generated based on a fixed probability. Each edge is then randomly assigned with a time lag between 0 and the max time lag τ_{max}, transforming the DAGs from ER model into T-DAGs. The k cluster centers are generated sequentially such that the T-SHD ($w_t = 1$ by default) between each pair of cluster center T-DAGs needs to exceed a distance lower-threshold $d_{between-min}$. The list of T-DAG members within each cluster is created by randomly flipping or removing d_{within} number of edges of the center T-DAG, such that the distance between cluster center and member is d_{within}. In summary, the key hyper-parameters of the simulated T-DAG clusters include graph parameters such as node number M and max time lag τ_{max}, as well as clustering structure parameters such as number of clusters k, within-cluster distance d_{within} and minimum threshold for between-cluster distance $d_{between-min}$. Pseudo code of T-DAGs generation is in Algorithm 2.

In this study, we test the robustness of the proposed clustering algorithm by varying cluster number k, max time lag τ_{max}, within-cluster distance d_{within} and node number M, corresponding to four separate experiments. When generating the simulated T-DAGs, the number of T-DAGs per cluster is 50. Since the true number of clusters is not provided to the clustering procedure, the optimal cluster number and weight w_t for T-SHD are selected by maximizing Silhouette scores during clustering. Candidate lists for weight w_t in T-SHD and cluster number are $\{0, 0.25, 0.5, 0.75, 1.0\}$ and $\{2, 3, \cdots, 14, 15\}$, respectively. Agglomerative clustering, a commonly used hierarchical clustering algorithm, is applied to obtain the cluster labels. After performing the clustering, the predicted cluster labels are evaluated using ARI against the true underlying cluster structure. The results of the four experiments are shown in Figure 2.

The results in Figure 2 show that the proposed clustering algorithm is robust to various cluster and graph settings. In most settings, the ARI of the proposed algorithm is above 0.85, except in extreme cases of within-cluster distance d_{within} and node number M. In Figure 2c, when d_{within} reaches 32, the average distance from center mem-

Algorithm 2: Random T-DAG Generation

Input: Node number M, Edge number M_{edge}, Max time lag τ_{max}, Cluster number k, Graph number list $\{N_j\}_j^k$, Distance within center d_{within}, Minimum distance between center $d_{between-min}$, Random ER graph generator $ER(M, M_{edge})$, empty list $l_{\bar{G}}$ for cluster centers, empty list l_G for T-DAGs,, T-SHD weight w_t (optional, default is $w_t = 1$)
Generate a T-DAG $\bar{G}_1 \sim ER(M, M_{edge})$ from ER graph and save it to $l_{\bar{G}}$
Generate N_1 T-DAGs by randomly remove or alternate d_{within} edges from $\bar{G}_1$, then save them to l_G
for $i \in \{2, \cdots, k\}$ **do**
    Repeatly generate $\bar{G}_i \sim ER(M, M_{edge})$ until it satisfies $\forall \bar{G}_j \in l_{\bar{G}}, T\text{-}SHD(\bar{G}_i, \bar{G}_j, w_t) \leq d_{between-min}$
    Add $\bar{G}_i$ to $l_{\bar{G}}$
    Generate N_i T-DAGs by randomly remove or alternate d_{within} edges from $\bar{G}_i$, then add them to l_G
end for
return The list of T-DAGs l_G

ber to cluster center is nearly half of the average distance between cluster centers. This extreme case of d_{within} is inherently difficult for clustering, as boundaries of clusters overlap with each other. This explains why ARI drops significantly at large d_{within} in Figure 2c. Similarly, in Figure 5d of Appendix B, when node number is smaller than 8, it reduces the average distance between centers, leading to the same effect as extreme d_{within} in Figure 2c. Detailed discussion is in Appendix B. Apart from clustering accuracy, the cluster number selected by the Silhouette score is always the same as the true cluster number, except for the two extreme cases discussed above.

Simulation study 2: Performance of Clustering on Estimated T-DAGs

In the second simulation study, we evaluated the clustering results on estimated causal graphs recovered from simulated longitudinal data. The clusters of T-DAGs are generated following the same procedure described in the previous section. Then each T-DAG is randomly assigned with positive or negative edge weights with magnitudes from $(0.5, 2.0)$. Longitudinal data is then generated corresponding to each T-DAG, where the time horizon is the same for the datasets of all T-DAGs. The variables are generated according to the edges in T-DAG, following a linear structural equation model (SEM) with Gaussian noises. Rather than providing the true T-DAGs to the clustering algorithm, we first apply PCMCI+ to recover an estimated T-DAG, as described in Algorithm 1. Then we use the agglomerative clustering algorithm on estimated causal graphs to return the cluster labels.

In this experiment, we mainly investigate whether CLOUD-CG is robust to the noises caused during the estimation of the true T-DAGs. We vary the cluster number of underlying clusters k, max time lag τ_{max} of T-DAGs, and time horizon of generated data for each T-DAG. Apart from the ARI score for clustering accuracy, we also report the normalized SHD for the causal discovery process. Normalized SHD measures the distance between the T-DAG estimated from the longitudinal data and the true underlying T-DAG. A normalized SHD close to 0 is desirable. The experiment results of clustering performance and causal discovery accuracy are in Figure 3.

Results of Figure 3 show that clustering results of CLOUD-CG achieve an ARI score above 0.8 in most cases. In the extreme case of $\tau_{max} = 2$ in Figure 3b, the normalized SHD is above 0.053, indicating 16 edges are falsely learned in the causal discovery procedure on average. The total expected edge number for all graphs is around 30, such that edges with learning error are more than half of the true edges in true underlying graphs in this extreme case. Such extreme causal graph learning errors are large enough to affect the boundary of clusters and significantly increase the difficulty of clustering. In all other settings, despite that causal discovery doesn't perfectly recover the true T-DAGs, our proposed Algorithm 1 is still robust to such causal discovery errors by achieving high ARI score and always selecting the correct cluster number.

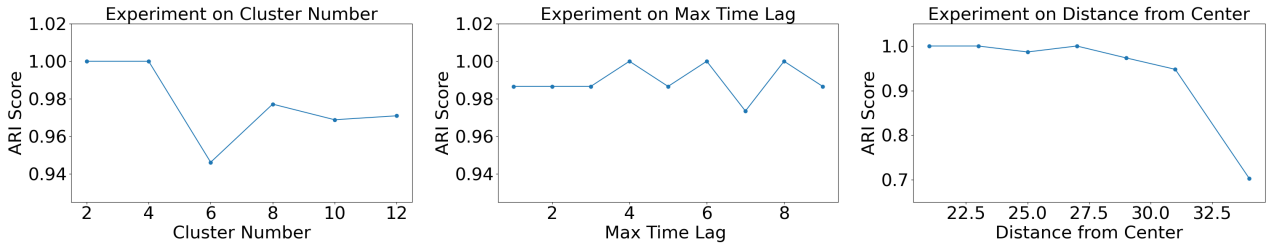
Application of CLOUD-CD to Deposit Data of Mexican Commercial Banks

Data and Preprocessing

We applied our methodology to 49 commercial banks in the Mexican market to identify the macroeconomic variables that most significantly influence their deposits. The specific macroeconomic variables considered are detailed in Table 1 of Appendix C. These macroeconomic variables were selected because they are the ones considered by the Central Bank of Mexico during stress testing (Banxico 2024). Additionally, we categorized deposits based on the type of counterparty and the deposit term (short-term or long-term), as outlined in Table 2 in Appendix C. Our analysis primarily focuses on demand deposits from natural persons as Diamond and Dybvig (1983) explains how banks provide liquidity to individuals through demand deposits and highlights why this structure makes banks vulnerable to bank runs. The dataset records monthly data points, covering the period from January 2006 to July 2024. Before conducting the analysis, we performed several data preprocessing steps, including adjusting for inflation and transforming relevant time series variables to achieve stationarity, for which the details are in Appendix C.

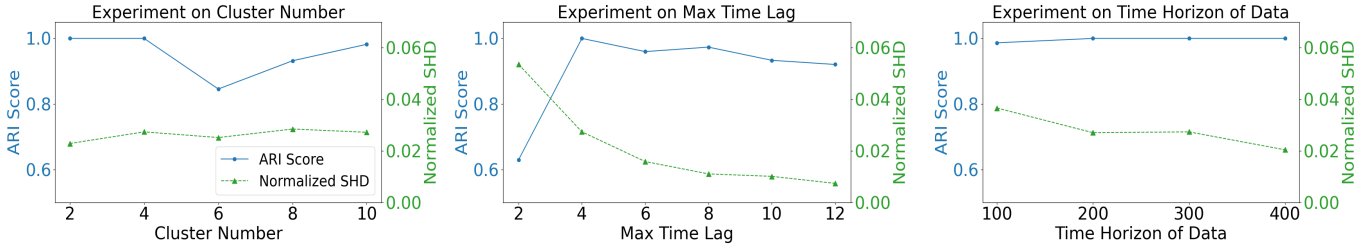
Motivation

While previous studies on the financial health of commercial banks in emerging markets focus on country-level panel data, we believe it is not sufficient to monitor the entire financial activity in the system. Specifically, such aggregated data fails to reveal interbank transactions, as those do not alter the overall total but significantly impact individual bank-level deposits. For instance, Figure 11 presents the aggregated PCMCI+ results at the system level. Notably, a single edge is observed pointing to individual deposits (`min_vista`), originating from the USDMXN exchange



(a) Experiment on cluster numbers (b) Experiment on time lags (c) Experiment on distances within cluster

Figure 2: Results of clustering on simulated T-DAGs.



(a) Experiment on cluster numbers (b) Experiment on time lags (c) Experiment on the horizon of data

Figure 3: Results of clustering accuracy on estimated causal graphs recovered from simulated data. Clustering accuracy (ARI score) is shown in a blue straight line against the scale on the left vertical. Causal discovery accuracy (normalized SHD) is shown in the green dashed line against the scale on the right vertical axis.

rate (`Tipo_Cambio`). This relationship is further corroborated by the regression results in Table 3 in Appendix D.

However, attributing deposits solely to the exchange rate would be an oversimplification. In Figure 12, we present the T-DAG for deposits across seven major banks after controlling for relevant macroeconomic variables. Notably, several positive edges are observed between these banks. A key factor to consider here is the extent of customer overlap. When there are substantial overlaps, a customer receiving funds in Bank A may transfer those funds to Bank B to pay credit card bills or loans. This interbank activity underscores the interconnectedness of the financial system and explains the observed positive relationships between deposits across banks. Analyzing only aggregated data obscures these interbank transactions, emphasizing the need to examine deposit patterns at the institutional level for a more comprehensive understanding.

Implementation Details

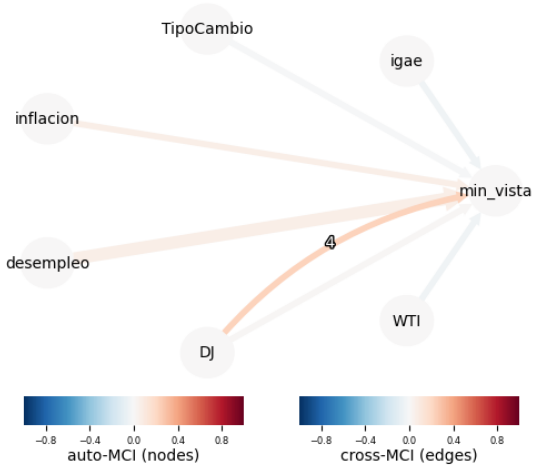
As described in the previous section, edges between units are significant, and we opted for a setting that accounts for inter-unit relationships. Furthermore, it is reasonable to assume that the direction of causality flows from the macroeconomic variable to the deposit, as the reverse direction of causal impact is significantly less likely. To account for this, we incorporate this prior knowledge into the discovery algorithm, ensuring that individual bank deposit data do not influence macroeconomic variables.

Before clustering, we tune the number of clusters k and the weight parameter w_t using the Silhouette score as the evaluation metric. As illustrated in Figure 9 in Appendix, the maximum Silhouette score is achieved at $k = 4$ and $w_t = 0.25$. Additionally, we applied T-DAG filtering to include only the nodes that ultimately lead to the natural person's deposit. For additional details on the implementation, including the choice of maximum time lag and relevant figures, please refer to Appendix E.

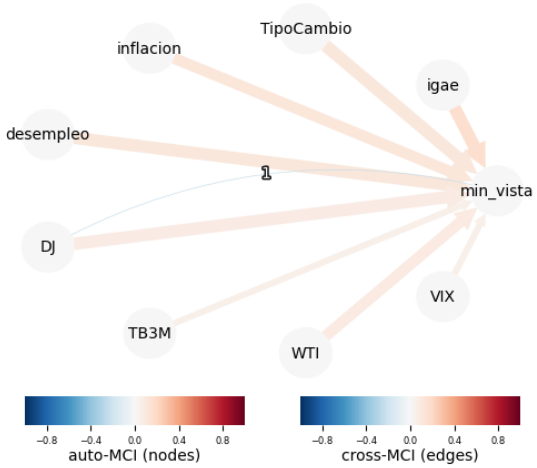
Results

We use the calculated T-SHD matrix with $k = 4$, $w_t = 0.25$, and $w_n = 1$. The results of the clustering can be seen in Figures 4. Figure 4 is a simplified aggregated T-DAG for institutions that belong to this cluster. If two or more institutions have the same edge, the edge weight is averaged and the width indicates the number of institutions sharing the edge. To maintain confidentiality, we have omitted the names of the banks involved. We have provided a distribution of bank characteristics for each cluster, highlighting their headquarters' locations and business models, as shown in Figure 13 of Appendix.

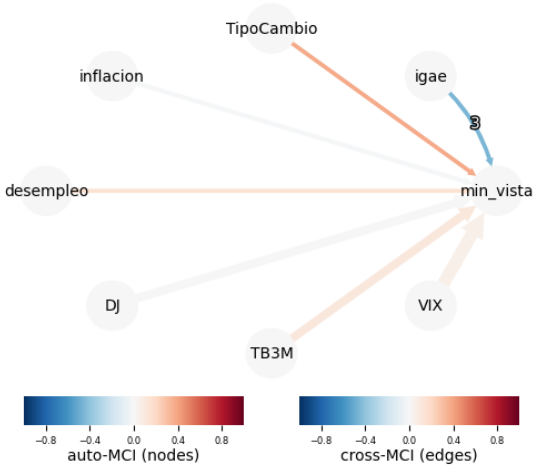
Our findings are summarized as follow: the drivers for deposits (node min_vista in the graph) in Cluster 0 predominantly consists of Domestic Systemically Important Banks, with unemployment (node Desempleo in the graph) and the Dow Jones (DJ) Index at a four-month lag identified as the primary influencing factors. Cluster 1 includes all North American-based banks among those four clusters, alongside several large national banks with greater asset sizes compared to those in Clusters 2 and 3. This cluster appears to be more sensitive to shocks from multiple macroeconomic



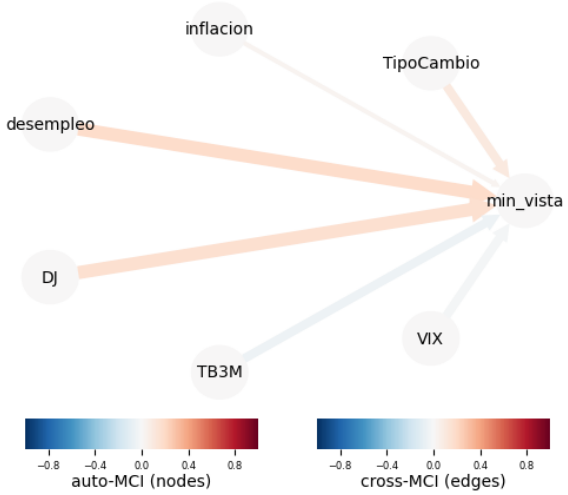
(a) Results for Cluster 0



(b) Results for Cluster 1



(c) Results for Cluster 2



(d) Results for Cluster 3

Figure 4: Combined Results for Clusters 0, 1, 2, and 3

variables, reflecting its central role in the financial system. Cluster 2 comprises smaller banks in terms of assets, where the Volatility Index (VIX) emerges as the dominant driver for deposits. Lastly, Cluster 3 consists of national commercial banks with a significant proportion of U.S. equities in their portfolios, where unemployment and the DJ Index with no time lag are identified as the strongest influencing factors for natural persons' deposits. For additional details on the results, please refer to Appendix F.

Interestingly, banks without causal edges from macroeconomic variables to deposits, which are less prone to shocks, are predominantly non-national institutions, with only 26.7% being national banks. This observation highlights how our clustering method effectively captures variations in resilience driven by differences in business models and exposure to macroeconomic factors. It is important to emphasize that while causal discovery algorithms provide a valuable framework for generating potential hypotheses and gaining insights into underlying relationships, their findings should be further validated through additional studies to ensure robustness and reliability.

Future Work

Future work for CLOUD-CG may include two directions. First, the algorithm components of our framework can be enhanced by incorporating more recent causal discovery and clustering algorithms. Second, CLOUD-CG's clustering labels can also serve as proxies for some hidden categorical variables, potentially improving downstream tasks such as causal inference.

References

Ali, G. B.; Bui, D. S.; Lodge, C. J.; Waidyatillake, N. T.; Perret, J. L.; Sun, C.; Walters, E. H.; Abramson, M. J.; Lowe, A. J.; and Dharmage, S. C. 2021. Infant body mass index trajectories and asthma and lung function. *Journal of Allergy and Clinical Immunology*, 148(3): 763–770.

Banxico. 2024. Reporte de Estabilidad Financiera, Junio 2024. Accessed: 2024-11-19.

Beckers, L.-M.; Brack, W.; Dann, J. P.; Krauss, M.; Müller, E.; and Schulze, T. 2020. Unraveling longitudinal pollution patterns of organic micropollutants in a river by non-target screening and cluster analysis. *Science of the Total Environment*, 727: 138388.

Birkenbihl, C.; Ahmad, A.; Massat, N. J.; Raschka, T.; Avbersek, A.; Downey, P.; Armstrong, M.; and Fröhlich, H. 2023. Artificial intelligence-based clustering and characterization of Parkinson's disease trajectories. *Scientific Reports*, 13(1): 2897.

Brouwer, E. D.; Arany, A.; Simm, J.; and Moreau, Y. 2021. Latent Convergent Cross Mapping. In *International Conference on Learning Representations*.

Cheng, Y.; Li, L.; Xiao, T.; Li, Z.; Suo, J.; He, K.; and Dai, Q. 2024. CUTS+: High-Dimensional Causal Discovery from Irregular Time-Series. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(10): 11525–11533.

Cheung, Y.-W.; and Lai, K. S. 1995. Lag order and critical values of the augmented Dickey–Fuller test. *Journal of Business & Economic Statistics*, 13(3): 277–280.

Chiou, J.-M.; and Li, P.-L. 2007. Functional clustering and identifying substructures of longitudinal data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(5): 679–699.

De Bock, R.; and Demyanets, A. 2012. Bank Asset Quality in Emerging Markets: Determinants and Spillovers. IMF Working Papers 2012/071, International Monetary Fund.

Den Teuling, N.; Pauws, S.; and van den Heuvel, E. 2021. Clustering of longitudinal data: A tutorial on a variety of approaches. *arXiv preprint arXiv:2111.05469*, 37.

Diamond, D. W.; and Dybvig, P. H. 1983. Bank Runs, Deposit Insurance, and Liquidity. *Journal of Political Economy*, 91(3): 401–419.

Gong, W.; Jennings, J.; Zhang, C.; and Pawlowski, N. 2022. Rhino: Deep Causal Temporal Relationship Learning with History-dependent Noise. In *The Eleventh International Conference on Learning Representations*.

Granger, C. W. J. 1969. Investigating Causal Relations by Econometric Models and Cross-Spectral Methods. *Econometrica*, 37(3): 424–438. Accessed 5 Nov. 2024.

Ha, S. H.; Bae, S. M.; and Park, S. C. 2002. Customer's time-variant purchase behavior and corresponding marketing strategies: an online retailer's case. *Computers & Industrial Engineering*, 43(4): 801–820.

Huang, D. Y.; Evans, E.; Hara, M.; Weiss, R. E.; and Hser, Y.-I. 2011. Employment trajectories: Exploring gender differences and impacts of drug use. *Journal of vocational behavior*, 79(1): 277–289.

James, G. M.; and Sugar, C. A. 2003. Clustering for Sparsely Sampled Functional Data. *Journal of the American Statistical Association*, 98(462): 397–408.

Kohlscheen, E.; Murcia Pabón, A.; and Contreras, J. 2018. Determinants of Bank Profitability in Emerging Markets. Technical report, BIS Working Paper No. 686. Available at SSRN: https://ssrn.com/abstract=3098196.

Lloyd, S. 1982. Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28(2): 129–137.

Lu, Z. 2024. Clustering Longitudinal Data: A Review of Methods and Software Packages. *International Statistical Review*.

McLachlan, G.; and Peel, D. 2004. *Finite Mixture Models*. John Wiley & Sons.

Mody, A.; Ohnsorge, F. L.; and Sandri, D. 2012. Precautionary Savings in the Great Recession. IMF Working Paper WP/12/42, International Monetary Fund.

Nagin, D. S. 1999. Analyzing developmental trajectories: A semiparametric, group-based approach. *Psychological Methods*, 4(2): 139–157.

Proust-Lima, C.; Philipps, V.; and Liquet, B. 2017. Estimation of Extended Mixed Models Using Latent Classes and Latent Processes: The R Package lcmm. *Journal of Statistical Software*, 78(2): 1–56.

Ren, R.; Fang, K.; Zhang, Q.; and Wang, X. 2023. Multivariate functional data clustering using adaptive density peak detection. *Statistics in Medicine*, 42(10): 1565–1582.

Rousseeuw, P. J. 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20: 53–65.

Runge, J. 2020. Discovering contemporaneous and lagged causal relations in autocorrelated nonlinear time series datasets. In Peters, J.; and Sontag, D., eds., *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI)*, volume 124 of *Proceedings of Machine Learning Research*, 1388–1397. PMLR.

Spirtes, P.; Glymour, C.; and Scheines, R. 2000. *Causation, Prediction, and Search*. Boston, MA: MIT Press.

Stepanyan, V.; and Guo, K. 2011. Determinants of Bank Credit in Emerging Market Economies. IMF Working Papers 2011/051, International Monetary Fund.

Sun, X.; Schulte, O.; Liu, G.; and Poupart, P. 2023. NTS-NOTEARS: Learning Nonparametric DBNs With Prior Knowledge. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, 1942–1964. PMLR.

Tank, A.; Covert, I.; Foti, N.; Shojaie, A.; and Fox, E. B. 2022. Neural Granger Causality. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(8): 4267–4279.

Thorndike, R. L. 1953. Who belongs in the family? *Psychometrika*, 18(4): 267–276.

WorldBank. 2024. Remittances Slowed in 2023, Expected to Grow Faster in 2024. https://www.worldbank.org/en/news/press-release/2024/06/26/remittances-slowed-in-2023-expected-to-grow-faster-in-2024. Accessed: 2024-11-14.

Zheng, X.; Aragam, B.; Ravikumar, P. K.; and Xing, E. P. 2018. Dags with no tears: Continuous optimization for structure learning. *Advances in neural information processing systems*, 31.

Zhou, J.; Zhang, Y.; and Tu, W. 2023. clusterMLD: An Efficient Hierarchical Clustering Method for Multivariate Longitudinal Data. *Journal of Computational and Graphical Statistics*, 32(3): 1131–1144.

Appendix A: Related Work

Time Series Causal Discovery (TSCD) Previous work on causal discovery methods includes Granger-causality-based methods (Granger 1969; Tank et al. 2022; Cheng et al. 2024), constraint-based methods (Runge 2020; Sun et al. 2023; Gong et al. 2022), and state space-based methods (Brouwer et al. 2021). In this work, we focus on (Runge 2020) because it provides detailed information on the number of lags, a critical factor for monitoring financial stability.

Longitudinal Data Clustering A recent paper reviews clustering methods for longitudinal data (Lu 2024). Model-based methods (Nagin 1999; Proust-Lima, Philipps, and Liquet 2017) are based on finite mixture models (McLachlan and Peel 2004). Algorithm-based approaches include K-means (Lloyd 1982) and hierarchical clustering (Zhou, Zhang, and Tu 2023), using the trajectory function. Functional clustering approaches (James and Sugar 2003; Chiou and Li 2007; Ren et al. 2023) use functional data analysis (FDA), where data are treated as functions from an infinite-dimensional space, unlike longitudinal data analysis, which views data as individual points. To the best of our knowledge, this is the first work on clustering longitudinal data using T-DAGs to identify sets of units based on how outcomes are affected by multiple covariates in a lag-specific manner.

Determinants of Deposits in Emerging Markets This work also contributes to the literature on the determinants of institution-level deposit outflows in emerging markets. The existing literature is relatively limited, largely due to confidentiality in commercial bank deposit data. Kohlscheen, Murcia Pabón, and Contreras (2018) analyzed bank profitability across 19 emerging market economies, concluding that higher long-term interest rates enhance profitability. De Bock and Demyanets (2012) find that macroeconomic factors significantly impact bank asset quality in emerging markets. Stepanyan and Guo (2011) examine the determinants of bank credit in these economies, concluding that strong economic fundamentals, favorable external conditions, and accommodative domestic monetary policy drive credit growth. While these studies provide valuable insights, their reliance on country-level panel data often obscures critical variations among individual banks. Our findings underscore the heterogeneity at the institutional level, highlighting the importance of analyzing granular deposit flows. Such detailed analysis can equip policymakers with the necessary information to make informed decisions and effectively maintain financial stability.

Appendix B: Discussion of Simulation Results on Extreme Cluster Settings

The results in Figure 5 show that the proposed clustering algorithm is robust to various cluster and graph settings. In most settings, the ARI of the proposed algorithm is above 0.85, except in two extreme cases in Figure 5c and 5d. In particular, in the experiment on within-cluster distance d_{within}, the average distance between cluster centers is around 60 in terms of T-SHD with $w_t = 1$ for all seven settings in Figure 5c. In general, enlarging d_{within} increases the difficulty of clustering. When d_{within} reaches 32, ARI drops below 0.75. In this case, our proposed clustering algorithm selects $w_t = 0$ in T-SHD to minimize the distances caused by the time lag difference between member T-DAGs within a cluster. Still, this cluster structure is inherently challenging for all clustering methods, as the distance between the center member and the cluster center is larger than half of the average distance between cluster centers, causing the boundaries of clusters to overlap. This explains why ARI drops significantly when d_{within} is beyond 32. Similarly, in Figure 5d, the ARI around 0.75 for node number $M = 8$ is also due to such a high intra-cluster distance structure. When node number M decreases, the space of all possible T-DAGs, $\mathbb{N}^{\mathcal{M}\times\mathcal{M}\times\mathcal{T}}$ shrinks, causing cluster centers to be closer to each other. In the experiment on node numbers, $d_{within} = 22$ for all settings. When $M = 8$, d_{within} is roughly half of the average distance between cluster centers, causing ambiguous cluster boundaries. Apart from clustering accuracy, the cluster number selected by the Silhouette score is always the same as the true cluster number, except for the two extreme cases discussed above.

Appendix C: Data Description and Preprocessing

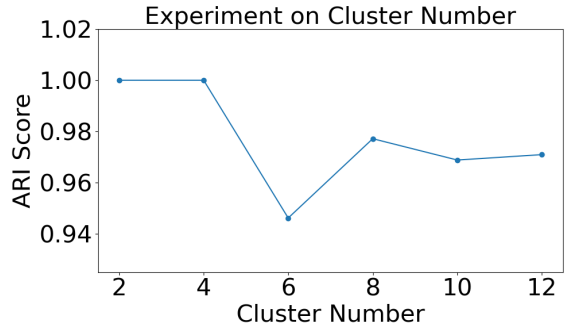
We utilize deposit data (Table 2), commercial bank loans (`vig_real`) provided by Banco de México (the Central Bank of Mexico), and macroeconomic variables (Table 1). While most macroeconomic variables are publicly available, the total real outstanding balance of commercial bank loans is not. The selection of macroeconomic variables is informed by their use in stress testing conducted by Banco de México.

To adjust nominal deposit values to real terms, we use the Índice Nacional de Precios al Consumidor (INPC), Mexico's official Consumer Price Index (CPI). This adjustment standardizes all nominal values by converting them to a comparable base period. Specifically, we transform each deposit using the following formula:

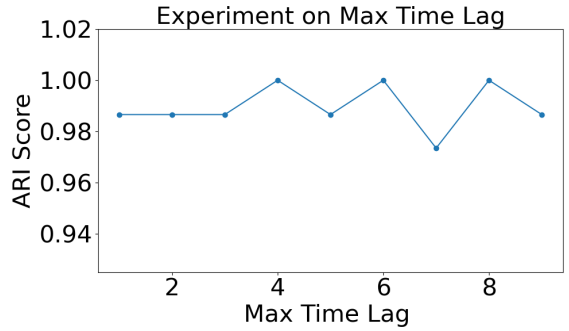
$$D_t^* = D_t \frac{INPC_t}{INPC_T}$$

where D_t^* is the inflation-adjusted deposit value at time t, D_t is the nominal value of the deposit at time t, $INPC_t$ is the INPC at time t, and T represents the base period, here set to July 2024. After applying this transformation, all deposit values are expressed in terms of July 2024 purchasing power, making them directly comparable across different time points.

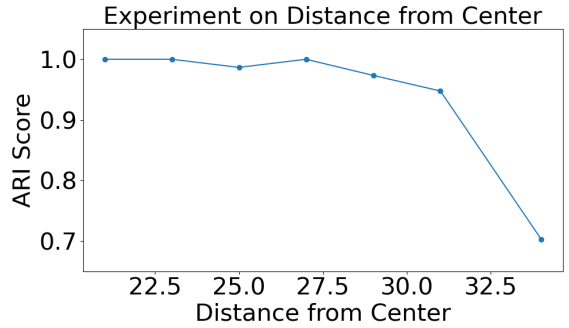
To meet the stationarity requirement for PCMCI+, we test each variable using the Augmented Dickey-Fuller (ADF) test (Cheung and Lai 1995). If the ADF test indicates non-stationarity, we first apply simple differencing and then reapply the ADF test. If the series remains non-stationary, we perform seasonal differencing with a time period of 12 months. The aggregated deposit value for the entire system before differencing is shown in Figure 6 and after differencing in Figure 7.



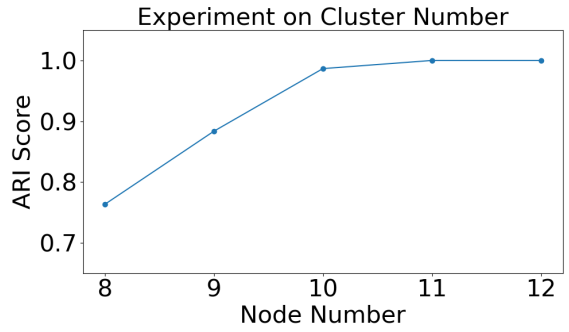
(a) Experiment on cluster numbers



(b) Experiment on time lags



(c) Experiment on distances between cluster component and cluster centers



(d) Experiment on node numbers

Figure 5: Results of clustering on simulated T-DAGs.

Variable	Description
IGAE	Global Indicator of Economic Activity measures Mexico's monthly economic activity.
TipoCambio	USD/MXN exchange rate.
Cetes28	Mexican 28-day treasury bill rate, an indicator of short-term interest rates.
Inflacion	Inflation rate measures the rate of increase in consumer prices.
Desempleo	Unemployment rate indicates the percentage of the unemployed labor force.
IPC	Mexican Stock Market Index (Índice de Precios y Cotizaciones), reflects the stock performance of large companies listed in Mexico.
DJ	Dow Jones Industrial Average, a US stock market index tracking 30 prominent companies.
TB3M	US 3-month Treasury Bill rate represents short-term interest rates in the US.
IPI_US	Industrial Production Index in the US measures the output of the US industrial sector.
WTI	West Texas Intermediate crude oil price is an indicator of oil prices.
TB10y	US 10-year Treasury bond yield represents long-term interest rates in the US.
VIX	Volatility Index measures the market's expectation of volatility (often associated with the S&P 500).
vig_real	Total real outstanding balance of commercial bank loans (including commercial, consumer, housing, and government loans).

Table 1: Description of Macroeconomic Variables

Variable	Description
ef_plazo	Term deposits from financial institutions.
ef_vista	Demand deposits from financial institutions.
gob_plazo	Term deposits from government entities.
gob_vista	Demand deposits from government entities.
min_plazo	Term deposits from individuals (natural persons).
min_vista	Demand deposits from individuals (natural persons).
ot_plazo	Term deposits from other entities, primarily legal entities (corporations).
ot_vista	Demand deposits from other entities, primarily legal entities (corporations).

Table 2: Descriptions of Deposits by Counterparty and Term

Appendix D: Results on aggregated deposit

Table 3 complements the results shown in Figure 11, indicating that the USDMXN exchange rate is the sole determinant

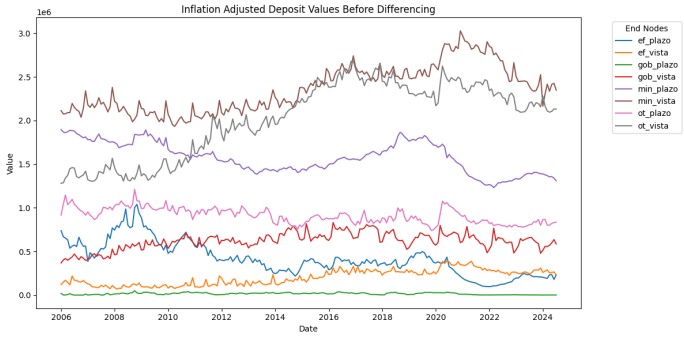


Figure 6: System Deposit Levels Before Differencing

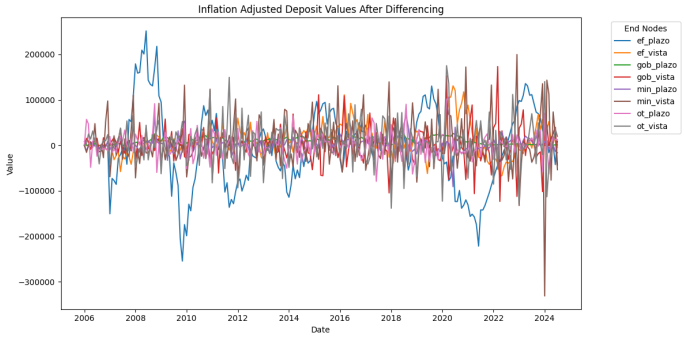


Figure 7: System Deposit Levels After Differencing

of system deposits. However, as discussed in the motivation section, this oversimplification may fail to capture the granular complexities of the financial system at an aggregated level, and basing decisions solely on this finding could lead to misleading conclusions.

Variable	**Coef.**	**Std. Err.**	**t**	$\mathbb{P} > \|t\|$
const	25170	28700	0.878	0.381
igae	-2208.93	2406.55	-0.918	0.360
TipoCambio	36770	12400	2.966	0.003***
Cetes28	1193.90	5861.55	0.204	0.839
inflacion	-1793.52	4144.50	-0.433	0.666
desempleo	-12950	22400	-0.577	0.564
IPC	-0.355	1.309	-0.271	0.786
DJ	13.678	8.938	1.530	0.127
TB3M	-1799.45	7708.46	-0.233	0.816
IPI_US	2451.00	2328.12	1.053	0.294
WTI	-283.91	294.06	-0.965	0.335
TB10y	-3812.70	9740.50	-0.391	0.696
VIX	315.16	870.24	0.362	0.718
vig_real	-0.0067	0.032	-0.208	0.836

Table 3: Regression Results for min_vista

There are several reasons why the exchange rate might influence individuals' deposits. One key factor is that many banks hold deposits in Mexican pesos (MXN) and U.S. dollars (USD), with MXN deposits being more prevalent. To calculate the total deposit value, the exchange rate is used to aggregate the amounts in both currencies, creating a direct dependency. Furthermore, Mexico is one of the largest recipients of remittances globally. According to the World Bank, Mexico received USD 66.2 billion in remittances in 2023, ranking second worldwide after India (WorldBank 2024). The volume of remittances is directly affected by fluctuations in the exchange rate.

Appendix E: Implementation Details

To determine the maximum time lag, $\tau_{\max}$, we performed a bivariate lagged conditional independence test for all pairs of covariates (8). Notably, the most significant lags occur within the first four months. Additionally, as shown in Figure 11, most lags are contemporaneous, which aligns with the nature of demand deposits as liquid assets, where the effects typically manifest quickly.

Appendix F: Discussion on Results for Application of CLOUD-CD to Deposit Data of Mexican Commercial Banks

Our analysis reveals that all clusters exhibit a positive contemporaneous relationship between unemployment (`desempleo`) and deposits. This observation aligns with the International Monetary Fund's findings that economic downturns lead to increased household saving rates (Mody, Ohnsorge, and Sandri 2012).

Regarding the Dow Jones Index (DJ), only Cluster 1 includes a North America-based bank, which shows a positive contemporaneous relationship within this cluster. Cluster 3 is directly influenced by DJ, with two commercial banks holding significant U.S. equity in their investment portfolios, thereby attracting more clients.

Cluster 2 is primarily driven by the Volatility Index (VIX). Notably, this cluster comprises eight relatively small banks, none of which are classified as D-SIBs (Domestic Systemically Important Banks). This suggests a potential research avenue: exploring whether unemployment impacts larger banks more significantly, while VIX affects smaller banks. In contrast, Cluster 0 contains three of the seven D-SIBs banks in Mexico.

This result reveals valuable insights into which banks may require closer supervision during economic downturns that impact specific macroeconomic variables. By identifying the relationships between macroeconomic indicators and individual banks, policymakers can proactively allocate supervisory resources to institutions that are more susceptible to particular economic conditions, thereby enhancing financial stability.

Notably, while over 60% of banks are national institutions, only 26.7% of banks without a causal relationship from macroeconomic variables to deposits are national 13. This finding suggests that banks headquartered outside Mexico may be more resilient to macroeconomic shocks than national banks. One contributing factor is that a significant portion of non-national banks conduct more business with legal entities. It is noteworthy that our clustering method effectively incorporates these characteristics into the results.

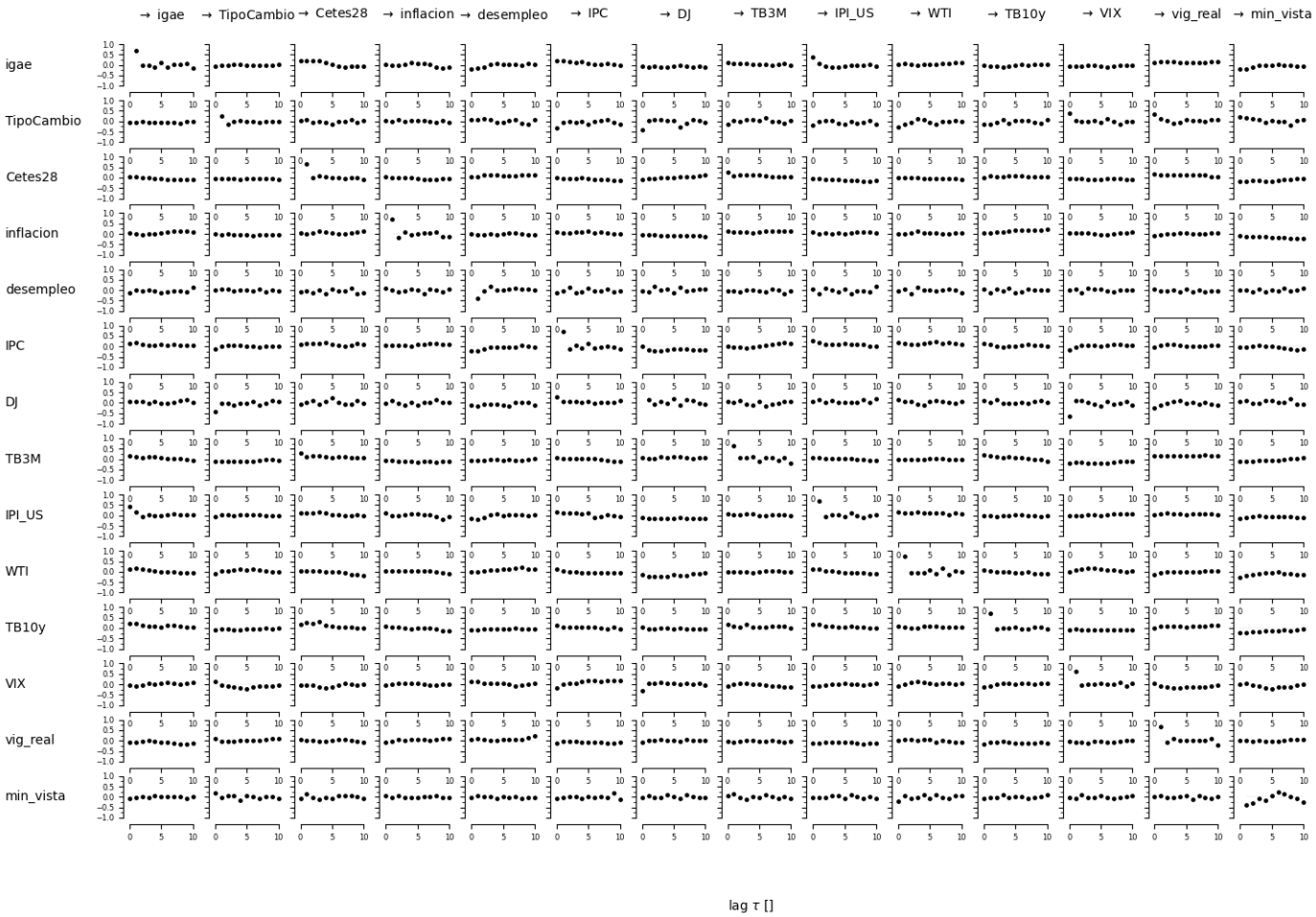


Figure 8: Bivariate, lagged conditional independence test

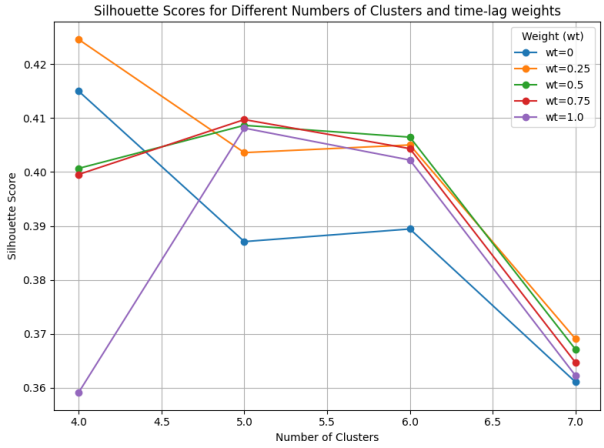


Figure 9: Tuning k and w_t based on Silhouette distance

Acknowledgment

We would like to thank Ron Bar, Samrat Bhattacharya, Liduvina Cisneros Ruíz, Jorge Luis García Ramírez, Uma Iyer, Fabrizio López Gallo Dey, Song Lu, Habib Saraya Jean, and Kaicheng Wu (listed in alphabetical order) for their valuable comments and suggestions, which greatly im-

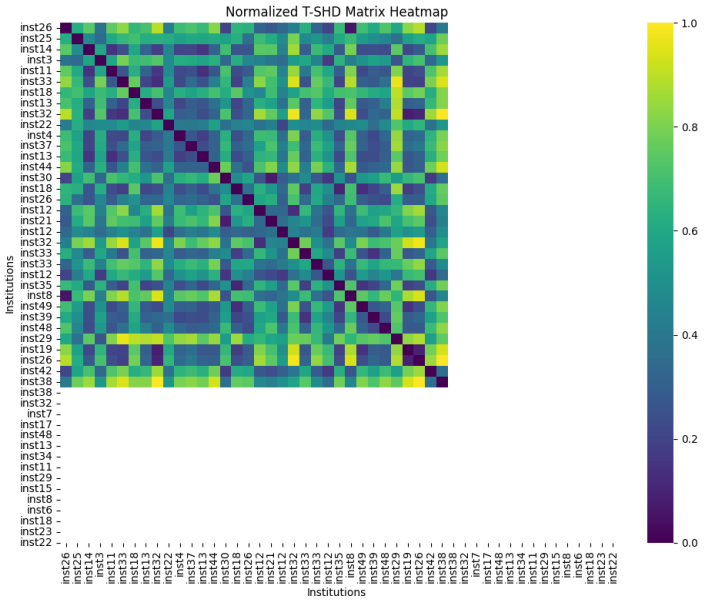


Figure 10: Normalized T-SHD Matrix Heatmap

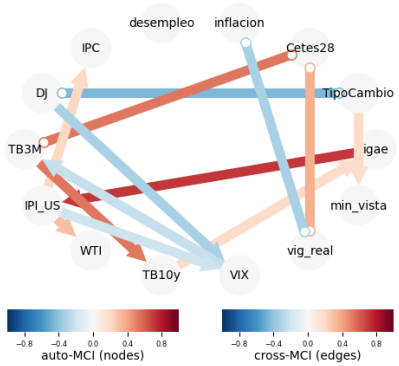


Figure 11: PCMCI+ Results on aggregated deposit

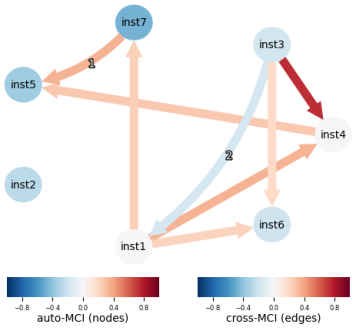


Figure 12: PCMCI+ Results between Major Bank Deposits

proved the quality of this work.

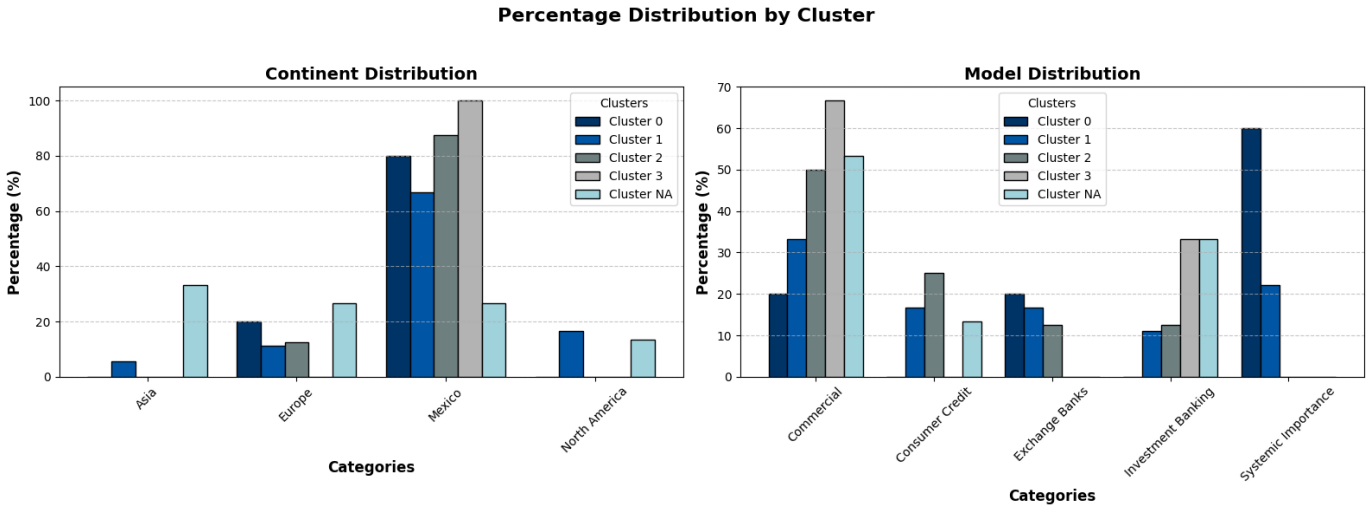


Figure 13: Distribution of Headquarters' Continents and Business Models Across Clusters