Discovering Migrant Types Through Cluster Analysis: Changes in the Mexico-U.S. Streams from 1970 to 2000

Extended Abstract - Do not cite or quote without permission.

Filiz Garip
Department of Sociology
Harvard University
fgarip@wjh.harvard.edu

September 4, 2008

Abstract

This paper explores types of migrants from Mexico to the United States in the period 1970-2000. Prior work analyzes the distinctions between migrants and non-migrants and suggests a number of theories that explain migration behavior. While each theory uncovers a different facet of migration flows, no single theory is able to capture the complexity of individuals' migration choices. Furthermore, focusing on what distinguishes migrants from non-migrants, prior research effectively treats migrants as a homogenous group, assuming that they respond to changes in the migration context in the same way. This paper develops a context-dependent model of migration and argues that variations in the social, economic and political context of sending and receiving regions create different conditions for migrating. These conditions are heightened or lessened by migrants' demographic characteristics and family networks. Hence, together all these elements help identify different types or strategies of migrants. A cluster analysis, informed by theories of migration, finds five distinct types of migrants from Mexico to the United States: network migrants (those who follow family or community migrants), income-maximizing migrants (those who seek to increase their income), risk-diversifying migrants (those who migrate to diversify their sources of income), push migrants (those who migrate to escape worsening economic conditions in Mexico), and pull migrants (those who take advantage of favorable migrating conditions to the U.S.). The relative presence and dominance of each migrant type follows a clear time pattern, signifying critical changes in the Mexican-U.S. migration context. Moreover, migrant types seem to influence several outcomes (legal or illegal entry, subsequent trips, length of stay), and lead to specific predictions not foreseen by the theories of migration. These results not only provide novel insights into the migration process between Mexico and the U.S., but they also show that different theories about why individuals migrate may each be correct in different contexts. Future research should focus on the interrelations among different theories of migration, and identify the specific contexts under which different ideas work.

1 Introduction

A variety of models reflecting different research objectives, focuses and interests have been proposed in the literature to explain international migration. Neoclassical economics has focused on wage and employment differentials between regions, conceiving of migration as an individual decision for income maximization (Todaro, 1969). The new economics of migration has viewed migration as a household decision to minimize risks to family income and to overcome capital constraints on family production activities (Stark et al., 1986). The segmented labor market theory linked migration to the structural requirements of modern industrial economics and viewed migration as created by a ‘pull’ created by labor demand (Piore, 1979).

More recent theories of international migration have suggested that while the conditions that initiate migration continue to cause people to move, new conditions may arise in the course of migration which come to function as independent causes themselves: migrant networks spread as suggested by the network theory, and create a feedback loop that makes migration more likely (Massey et al., 1993).

While each of these explanations offers some insight into the nature of international migration, no one explanation is satisfactory. Namely, these theories do not provide an understanding of the differences among *migrants*, since each theory seeks to only distinguish migrants from non-migrants. The empirical results typically show that migrants tend to be young, less educated, middle- or low-income individuals, who live in poor communities with few employment opportunities and have some connections to prior migrants through their social networks. This portrait of a ‘typical’ migrant, emerging from multivariate regression methods applied on large survey data sets, disregards the differences among migrants, and treats them as a homogenous group. An underlying assumption, then, is that migrants respond to changes in the migration context in the same way. None of the explanations in the literature allows us to observe the distinctions among migrants. By stark contrast, qualitative work in the literature has identified a variety of reasons and strategies underlying individuals' migration behavior. Yet, no quantitative study to date has questioned whether there may be different 'types' of migrants in different migration contexts.

This paper seeks to fill this gap in our understanding of the migration process. Specifically, the paper argues that individuals' migration behavior is heterogenous and may contain elements of different theories or scenarios. First, variations in the social, economic and political context of sending and receiving regions may create several different conditions for migrating. Second, these conditions may work differently for individuals depending on their demographic characteristics or social networks to prior migrants. Together, these elements help to identify different profiles or 'types' of migrants in specific contexts.

Therefore, we need to find out how the several factors which seem to affect individuals' migration behavior *interrelate* in different contexts. One way of analyzing the interdependency of different factors is using multivariate regression methods, as commonly done in the literature. Because the simple multivariate model assumes each factor to distinctly and independently affect migration, to analyze specific contexts, we need to introduce complex interactions of different factors. However, the number of possible interactions increases exponentially with the number of factors considered (e.g., for 3 variables, we need $2^3 = 8$ terms), making the model quickly unmanageable.

An alternative methodology is to use cluster analysis, and try to identify types of migrants with specific configurations of factors in the data. Cluster analysis is basically a search technique for locating groups of individuals who have similar scores on a series of variables. Not to be confused with factor or principal components analysis which reveal patterns across variables , clustering technique reveals patterns across cases (that is, migrant individuals).

Clustering can allow us to situate different theories of migration within specific contexts, and add precision to their hypotheses. The method can further help us discover unforeseen relationships among different factors within each group of migrants. Note that this approach, of focusing on the specific surrounding circumstances and coexisting conditions under which our findings hold, is similar to the philosophy of qualitative or small-scale quantitative case studies (Miller & Friesen, 1977).

In the remainder of the paper, we first review the prominent theories of migration, and explain the clustering methodology. Then, we use a large data set collected as part of the Mexican Migration Project, which records U.S. migration moves of 3,850 household heads in 114 Mexican communities from 1970 to 2000. We employ cluster analysis to identify five distinct types of migrants in the data, and use migration theories to interpret the unique conditions of each group. Then, we show how each migrant type is associated with different strategies (legal or illegal border crossing, length of stay, subsequent trips) and outcomes (naturalization in the U.S.). Finally, as this paper is only a preliminary draft, in the conclusion, we explain the future steps we will take to obtain additional findings and to tighten the arguments.

2 Background

Emergence of migration as a major force throughout the world has attracted considerable scholarly attention and has led to the development of a fragmented set of theories of international migration. Most of these theories posit causal mechanisms that are not inherently contradictory, and substantial effort has been made to construct a common framework using the complementarity of different ideas by Massey & Espinosa (1997) and Massey et al. (1994).

Neoclassical economics asserts that migration is caused by geographic differences in the supply and demand of labor. The resulting differentials in expected wages (Todaro, 1969) and employment conditions cause workers from the low-wage regions to move to the high-wage regions. As a result of this movement, the supply of labor decreases and the wages rise in the sending region, while the supply of labor increases and the wages fall in the receiving region. This perspective implicitly assumes that labor markets are the primary mechanism by which labor flows are induced and that elimination of wage differentials will end migratory movement.

According to the new economics of migration, a wage rate differential is not a necessary condition for migration to occur. Diversifying sources of household income in order to minimize risks is a viable reason to migrate as is maximizing income. Consequently, the new economics of migration relaxes the implicit assumption of neoclassical economics, which identifies labor markets as the primary mechanism underlying labor flows (Stark et al., 1986).

Although not in inherent conflict with neoclassical economics or the new economics of migration, the segmented labor market theory does carry corollaries that are not implied by either. According to this model, a wage rate differential is neither a necessary nor a sufficient condition for migration to occur, rather migration stems from the labor demand that grows out of the structural needs of the economy (Piore, 1979). Thus, being demand-based, the segmented labor market approach predicts that, at the community level, pull factors in the destination are stronger than push factors in the origin.

While these models present conditions for the initiation of migratory movement, empirical evidence suggests that conditions for its perpetuation may be quite different. Namely, theories of the perpetuation of migration suggest that acts of migration systematically change the context within which future migration decisions are made, and the structural changes thus created increase the likelihood of future migration. According to network theory, migration develops an increasingly dense web of contacts between sending and receiving regions that increases the likelihood of movement by lowering the costs and increasing the expected net returns to migration. Explicitly, network connections constitute a form of social capital that people can draw upon to gain access to employment in destination. This powerful mechanism -despite the strong and consistent evidence in support accumulated so far- is clearly undermined by theories of neoclassical economics, the new economics and the segmented labor market.

Each of these theories has received considerable support from the empirical studies in the literature. Namely, evidence suggests that each theory explains a different facet of an enormously complex subject matter and cannot be rejected on its own terms (Massey et al., 1994). Based on this evidence, it is clear that all models are 'correct', and either one by itself constitutes an incomplete explanation of migration. Therefore, a simultaneous examination of different theoretical propositions is necessary to achieve a full understanding of the dynamics of migration.

Researchers in the past have combined different models of migration by including a variety of indicators representing each theory, and testing their impact on migration using multivariate regression techniques (e.g. Massey & Espinosa, 1997). While a viable approach, this methodology required the assumption of independence among different theories. No study to date has looked at the interrelations among different theories of migration, nor identified the specific contexts under which each theory works.

This paper argues that unique configurations of the migration context and individuals' own or family characteristics create unique migrant types, strategies and outcomes. To identify these configurations, rather than multivariate regression techniques that focus on the similarity between variables, this paper suggests using cluster analysis, which focuses on the similarity between cases, namely migrant individuals. By using cluster analysis, we let data, not our own previous conceptions on migration behavior, determine the migrant groups. Then, once the clusters are determined, we use the existing theories of migration in their interpretation. This methodology provides insights beyond those afforded by regression methods alone, as we will see in the Preliminary Results section. But first, the following section explains the methodology in detail, and provides a description of the study setting.

3 Methodology

The premise of cluster analysis is that examining similarity between configurations of cases, rather than similarity between variables, may provide a useful perspective from which to understand migration behavior. Cluster analysis is a widely used technique for data classification in fields as diverse as computer science, biology and neuroscience. Social science applications of this methodology have become popular in recent years, predominantly in economics and sociology. (See Bailey (1975) for a review of the applications in sociology.)

Simply stated, cluster analysis is a method to group cases on the basis of their similarity on one or more measures. Then, to determine clusters in the data, we first need to select a measure of similarity between individuals, which would utilize data from several variables of interest. We select the correlation coefficient C as the measure of similarity, which is computed for each observation pair i and j as follows:

$$C = \frac{\sum_{k=1}^{p}(x_{ki} - \overline{x}_{\cdot i})(x_{kj} - \overline{x}_{\cdot j})}{\left\{\sum_{k=1}^{p}(x_{ki} - \overline{x}_{\cdot i})^2 \sum_{l=1}^{p}(x_{lj} - \overline{x}_{\cdot j})^2\right\}^{1/2}} \tag{1}$$

where x_{ki} is the value of variable k for observation i and p is the total number of variables.

As a clustering technique, due to its simplicity and computational speed, we select a means-based partitioning method. This method creates n clusters through an iterative process: each observation is assigned to a cluster whose mean is closest, and based on that classification, new cluster means are computed. The iterations continue until no observations change clusters. (Note that this method leads to partitions that are not hierarchically related. In future iterations of the paper, we will experiment with hierarchical clustering methods, and alternative measures of similarity.)

Since the number of clusters, n, are determined by the user, we need a stopping rule to determine the optimum number of clusters. The goal is to achieve the most distinct clustering in the data, that is most informative without being redundant. There are many stopping rules suggested in the literature. In this paper, we devised an add-hoc stopping rule to choose the cluster number which is described in a subsequent section.[1]

[1] Stata has only implemented the Calinski stopping rule for means-based partitioning. In the future, we will implement and try several other stopping rules in Matlab platform.

3.1 Study Setting and Data

We analyze the life history data collected from a random sample of 3,850 Mexican household heads in 114 communities as part of the Mexican Migration Project.[2] The survey data have been gathered in the winter months of 1982-2006 in communities located in western Mexico, a region that has historically been a major sender of migrants to the United States. These data have been supplemented with non-random samples of migrants located in the United States during the summer subsequent to each winter's survey. (For details of the data collection strategy, see Massey & Espinosa (1997).)

The life history data has been collected retrospectively from each household head in the sample. As only household heads are included, the data comes predominantly from men. The life histories include detailed information on migration and labor experiences of individuals, as well as property, marital and fertility information. These individual- and household-level data are supplemented with several community-level and macroeconomic indicators.

A data set that combines the individual, household and community level data with macroeconomic and policy context indicators, has been prepared and used by Massey & Espinosa (1997). The authors have made this data set publicly available. The analysis in this paper uses a similar data set with a larger number of communities. (Because their research was completed in 1997, they were able to include 25 communities. We include all 114 communities currently available.) Although the life history data goes as far back as 1900, the macroeconomic indicators of interest become available after 1970. Therefore we only include the 1970-2000 time period in our analysis. The life histories contain information on all migration moves by household heads from their birth until the survey year. Because the clustering methods we use cannot handle time-series data, we restrict our analysis to first-time migrants, and only keep the information from the year of their first trip. Besides, some migrants in the sample are born in the United States, or have migrated there in their childhood with their families. As we do not have any information on the nature of such moves (that is, we do not know anything about migrants' parents, or the context of migration then), we restrict our sample to only include the individuals who migrated after the age of 18.

[2] The Mexican Migration Project (MMP) is a collaborative research project based at the Princeton University and the University of Guadalajara" More information is available at http://mmp.opr.princeton.edu/

3.2 Variables and Operational Definitions

We use the variables defined (and made available in a data set) by Massey & Espinosa (1997) for analysis. The variables contain demographic and migration-related information about individuals, as well as indicators of the macroeconomic and policy context of Mexico and the United States, and allow us to account for the predictions of the migration theories reviewed earlier. For the clustering analysis, we only include variables that are measured on a continuous scale, and standardize them to have a mean of 0 and standard deviation of 1. This practice ensures that high variability or scale of a variable does not dominate the cluster analysis.(We leave out the binary indicators in Massey & Espinosa (1997) because the distance measures used in clustering can deal with either binary or continuous variables. There is a newly implemented distance measure in Stata, called Gower dissimilarity, which works with mixed data. We will experiment with this measure as a next step for the paper.)

Some of the operational measures the authors used have been modified to better capture the variable of interest. For example, labor market experience is measured as total years of labor market experience, rather than years since first job. In the latter definition, a person may be out of work for several years, but we would still count those years as part of the accumulated labor experience. Land is measured not as with a binary indicator, but with a continuous measure of hectares owned by the household. Similarly, businesses and properties are measured in numbers, rather than binary indicators. Table 1 lists all the variables and operational definitions used in the paper.

4 Preliminary Results

4.1 Selection of Clusters

We use the average dissimilarity among clusters as a stopping rule to determine the number of clusters. Namely, for each cluster solution, we compute the average dissimilarity among cluster pairs. For instance, for a 3-cluster solution, we take the average value of the distances between the first and second clusters, the first and third clusters, and the second and third clusters. To compute the dissimilarity between each cluster pair, we first compute the centroid of each cluster, which takes on the mean value of each variable over all the observations in the cluster. Then, for each cluster pair, we compute the correlations between the cluster centroids. Note that the correlation measure used in clustering captures similarity, not dissimilarity, and can take on negative and positive values. Since the direction of the relationship among clusters is of no importance, we first take the absolute value of

Table 1: Definition of Variables

Variable	**Operational Definition**
Demographic background:	
Age	Age at last birthday
No. of minors in household	No. of own children under age 18
General human capital:	
Labor market experience	Total number of years employed
Education	No. of year of school completed
General social capital:	
No. of U.S. migrants in family	No. of parents or siblings with U.S. experience
% U.S. migrants in community	Proportion over age 15 with U.S. experience
Physical capital:	
Land	Hectares of land owned by household
Home	No. of houses owned by household
Business	No. of businesses owned by household
Community economic context:	
% earning twice minimum wage	Proportion of workers earning at least twice the legal minimum wage
% self-employed	Proportion of workers who are self-employed
% females in manufacturing	Proportion of female workers in manufacturing
% of males in agriculture	Proportion of male labor force in agriculture
Proportion of arable land	Cultivable land divided by total land base
Macroeconomic context:	
Peso devaluation	Rate of change in dollar value of Mexican peso
Mexican inflation rate	Rate of change in Mexican consumer index
Growth in foreign investment	Rate of change in direct foreign investment
Mexican minimum wage	Minimum wage in Mexico (in 2000 U.S.$)
U.S. employment growth	Rate of change in total U.S. employment
U.S. average wage	Average wages in the U.S.(in 1990 U.S.$)
U.S. policy context:	
Availability of visas	Legal immigration divided by sum of legal and illegal entries
Probability of apprehension	Likelihood of arrest while attempting to cross border without documents

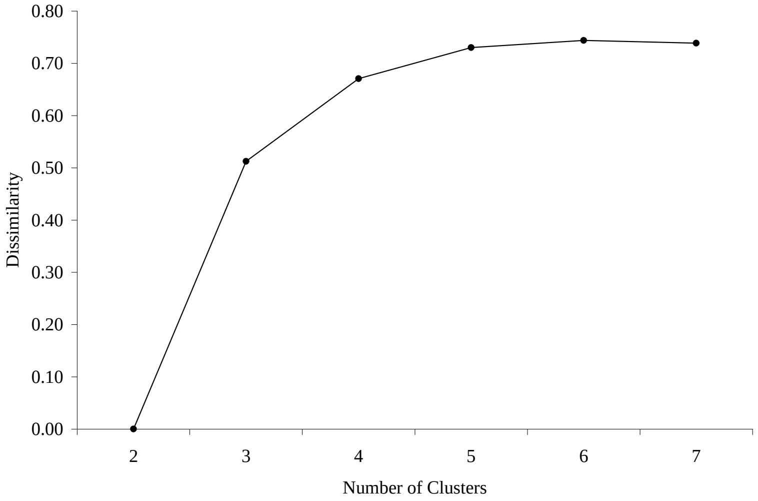

Figure 1: Average Dissimilarity among Clusters

the correlation measure, and then convert it to a dissimilarity measure by subtracting it from 1. Then, for each cluster solution, we average the values of dissimilarity between cluster pairs. As our goal is to obtain a more distinct clustering, higher values of the average dissimilarity are preferable.

The results are displayed in Figure 1. The horizontal axis shows the number of clusters the data is dissected into, and the vertical axis shows the average dissimilarity between the cluster pairs for that cluster solution. From this plot, one can see that the average dissimilarity among clusters increases with the number of clusters until the 5-cluster solution, after which it remains stable. We can then conclude that dissecting the data into 5-clusters generates the most distinct groupings in the data.

4.2 Interpretation of Clusters

Summary results from the cluster analysis are presented in Table 2, which lists the mean values of the variables for observations in each of the five clusters. (Note that we use the unstandardized mean scores for each cluster to better interpret the migrant profiles.) The pattern of variable means in each cluster, interpreted in light of the theories of migration presented earlier, provides a basis for understanding the nature of the cluster (Fleishman, 1986).

Cluster 1 contains the *risk-diversifying migrants*. Compared to the other groups, these people are older and tend to have more children. They have the lowest level of education among all clusters, yet the highest level of labor market experience. The migrants in this category are considered risk-diversifying because they own property and businesses, and they tend to migrate when the economic conditions in Mexico are not favorable. The average inflation rate for this cluster is 24.5% and the devaluation of Mexican peso is 0.21, both second-highest values among all clusters. Because these migrants are much older compared with other groups (48 years old as opposed to the overall sample average of 32), they are likely to be retired. Then, the unfavorable economic conditions in Mexico may be leading these individuals to migrate to United States to earn a living and to protect their investments at home. This type of migration is most closely related to the new economics of migration theory (Stark et al., 1986).

Those in cluster 2 are young (28.9 years old on average), have low education, low physical capital. They tend to live in communities where the majority of the male labor force is in agriculture (62%) and the proportion of arable land is high. Such communities are likely to be small, rural towns rather than metropolitan cities. (Note that the data set contains urban as well as rural communities). The characteristic that distinguishes this group from other clusters is the relative states of the Mexican and U.S. economies. Mexican wages take on their lowest value for this cluster (1203$/year), while, by stark contrast, the U.S. wages are at their highest value (11.25 $/hour). The wage differentials are the primary cause of migration according to the neoclassical theory (Todaro, 1969). Hence, we consider the migrants in cluster 2 to be acting in line with the expectations of this theory, and call them *income-maximizing migrants*.

Cluster 3 contains migrants whose distinguishing characteristics are high education (9.22 years on average) and high number of prior migrants in their families (2.15 family members). They tend to live in areas where a very low percentage of labor force is in agriculture (16%) and a high percentage of the population earns more than twice the minimum wage (43%). These areas are likely to be urban, and these migrants are likely to be following the prior migrants in their families. Note that the economic conditions in Mexico are not unfavorable for this group (the inflation and devaluation are both average, the minimum wage is relatively high, 2702 $/year), nor the wages or the policy context in the U.S. are very attractive compared with other groups (the average wage is only 10.12 $/hour, the visa accessibility is low). Given that there are no economic or political incentives for these individuals to migrate, we call them *network migrants* to suggest that they may be migrating to follow their family members in the U.S.

Members of cluster 4 are distinct in the U.S. policy context. Namely, they face the least set of political restrictions to cross the Mexican-U.S. border. The visa availability obtains its highest value compared with other clusters (0.08) and the probability of apprehension is at its lowest (0.27). Despite the robust position of the Mexican economic context for this cluster (below average inflation of 21%, low peso devaluation of 0.11, and high wages of 2865 $/hour), the favorable border policy seems to attract the migrants in this category. Accordingly, we call these migrants *pull migrants*.

Cluster 5 contains *push migrants*. The migrants in this category face the worst possible economic conditions in Mexico, with record-high inflation of 66% and peso devaluation of 1.30. Similarly, the minimum wage in Mexico is at the low value of 1826 $/year. Other than the economic conditions, the migrants in this cluster do not have any features that distinguish them from the migrants in other clusters. Hence, we call them *push migrants* to suggest the bad Mexican economy as the cause of their move to the United States.

The number of people in each cluster appears in the last row of Table 2. The distribution of migrants across the five clusters is relatively uniform. The income-maximizing cluster (#2) seems to contain a slightly higher number of migrants.

4.3 Principal Components Analysis

In order to assess the validity of clusters, namely evaluate whether clusters indeed identify distinct groupings in the data, Everitt et al. (2001) suggests using principal components analysis as a visual tool. Principal components analysis (PCA) aggregates the information from several variables (22 in our case) into a few dimensions. We can use these dimensions as graphical axes and plot the cluster membership of the observations. If cluster analysis indeed yields distinct groups in the data, then we should be able to observe distinct 'clouds' for each cluster on the graph.

We obtain the first three principal components, which explain 38% of the variation in the data (results available from the author upon request). We then use the cluster indicators to label the observations, and plot them against the PCA dimensions. Figure 2 displays the results. We can clearly distinguish separate clouds of observations for each cluster, and verify the validity of the groupings in the data suggested by the cluster analysis.

Table 2: Cluster Means

	Clusters				
Variable	1	2	3	4	5
Age	48.40	28.90	30.39	28.14	29.35
No. of minors in household	2.51	1.67	1.43	1.54	1.58
Labor market experience	34.94	13.50	15.29	14.03	15.30
Education	4.19	5.25	**9.22**	6.96	6.54
No. of U.S. migrant family members	1.45	1.36	**2.15**	1.59	1.73
% U.S. migrants in community	0.19	0.17	0.16	0.22	0.21
Land	0.15	0.29	0.04	0.14	0.12
Home	**1.04**	0.24	0.42	0.38	0.38
Business	**0.18**	0.05	0.07	0.07	0.07
% earning twice minimum wage	0.27	0.17	0.43	0.17	0.19
% self-employed	0.28	0.28	0.18	0.42	0.36
% females in manufacturing	0.23	0.22	0.30	0.19	0.22
% of males in agriculture	0.42	0.61	0.16	0.64	0.55
Proportion of arable land	0.59	0.79	0.65	0.78	0.77
Peso devaluation	0.21	0.08	0.16	0.11	**1.30**
Mexican inflation rate	24.50	16.04	21.72	21.15	**65.70**
Growth in foreign investment	0.24	0.21	0.32	0.42	-0.12
Mexican minimum wage	2528	**1203**	2703	2865	1826
U.S. employment growth	0.003	0.002	0.001	0.002	0.000
U.S. average wage	10.24	**11.25**	10.12	10.13	10.39
Availability of visas	0.05	0.06	0.07	**0.08**	0.02
Probability of apprehension	0.29	0.40	0.28	**0.27**	0.31
N	2392	4040	2727	2939	2968

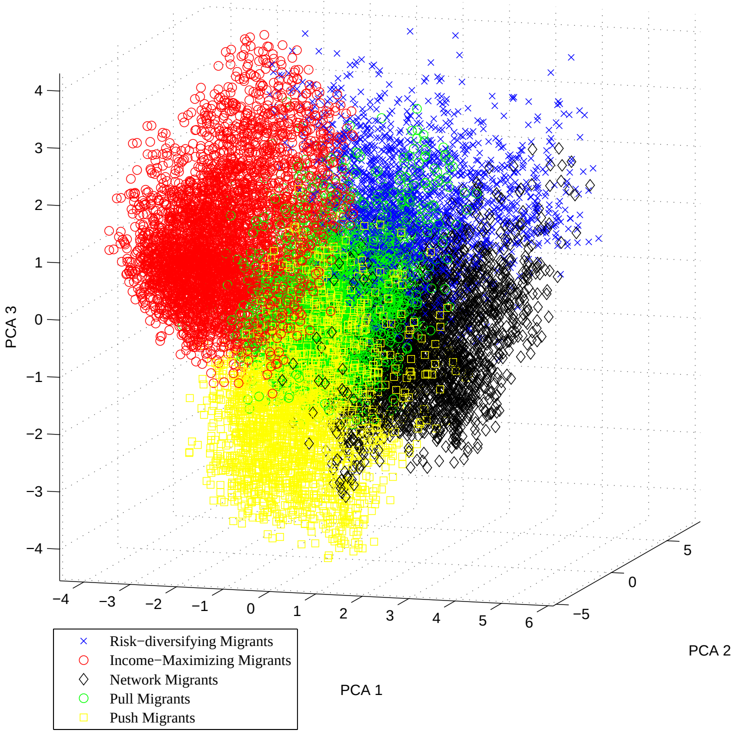

Figure 2: Configuration of Migrant Clusters across Three Principal Component Dimensions

4.4 Change in Migrant Types Over Time

Our data covers the migration flows between Mexico and the United States over a 30-year time period, from 1970 to 2000. Although cluster analysis utilized time-variant variables (such as inflation rate, minimum wage, etc.) to classify migrants, we have not yet considered how the composition of the migrant stream across the 5 types changes over time. Figure 3 plots the distribution of migrant types over time. The figure displays a striking time pattern to the presence and relative dominance of each migrant type over time. Namely, in the early years of the study period, migrants are predominantly income-maximizers. From 1980 to 1990, due to worsening economic conditions in Mexico, we observe an increasing number of ‘push’ migrants. Risk-diversifying and network migrants are observed almost each year, in increasing numbers over time. Pull migrants are concentrated mostly in the period after

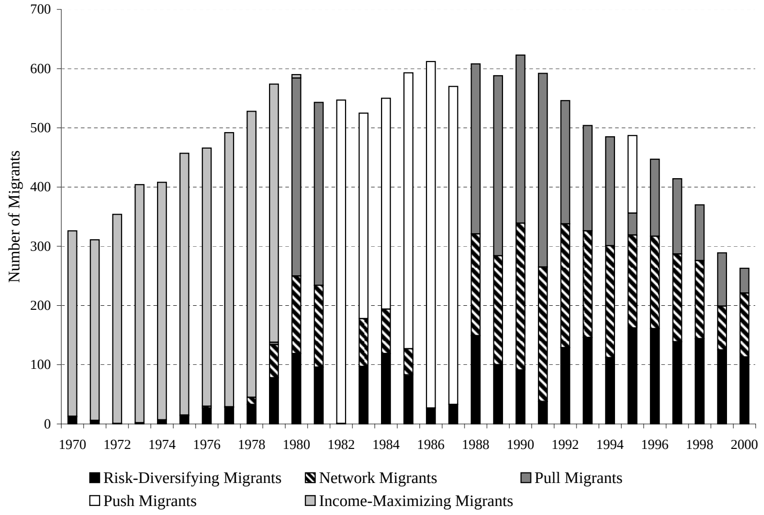

Figure 3: Distribution of Migrant Types Over Time 1970-2000

1988, and decrease in numbers with each year.

These patterns overlap with critical changes in the economic and political context of Mexico and the United States. Namely, as we see in Figure 4, for instance, the number of income maximizing migrants closely follows the U.S. average wage. Actually, in line with the expectations of the neoclassical theory, we observe no income-maximizing migrants after wages drop in the receiving region (possibly due an increasing availability of low-wage migrant workers).

Similarly, Figure 5 shows that the trend in the number of pull migrants over time is very close to the trend of U.S. visa availability. We observe increases in the number of pull migrants in our data as the trend-line for visa availability spikes. Finally, Figure 6 shows that the number of push migrants over time is closely related to the Mexican inflation rate, reaching high numbers only when inflation rate is at its peak. These plots actually provide an insight into how the cluster analysis has formed groups based on trends in macroeconomic indicators. While several different types of migrants may be present at the same time period, their relative dominance changes over time.

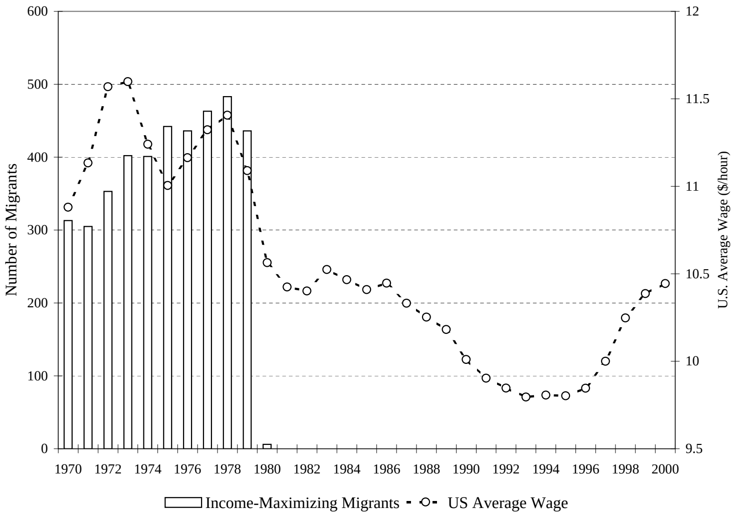

Figure 4: Number of Income-Maximizing Migrants and Mexican Inflation Rate Over Time 1970-2000

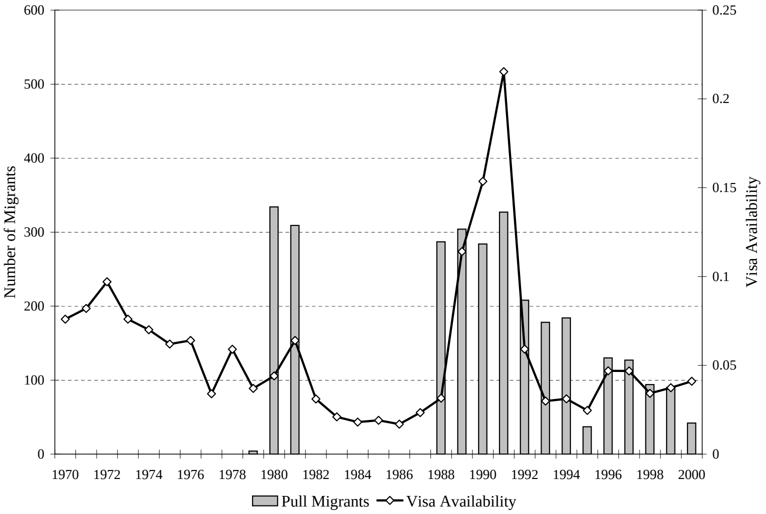

Figure 5: Number of Pull Migrants and U.S. Visa Availability Over Time 1970-2000

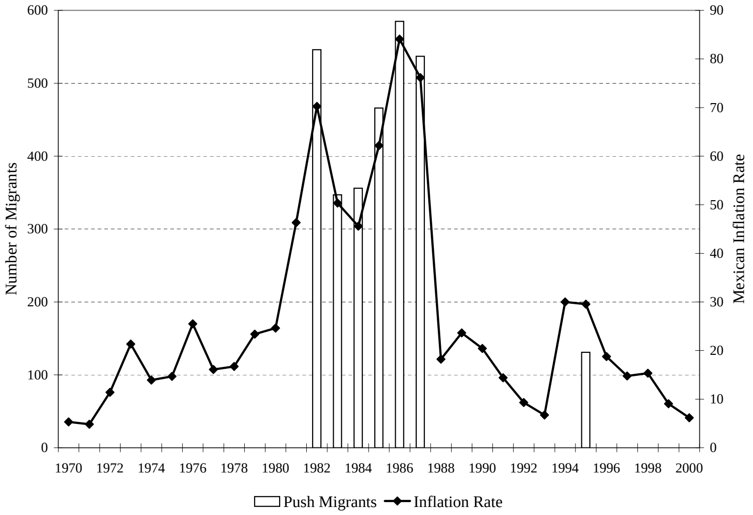

Figure 6: Number of Push Migrants and Mexican Inflation Rate Over Time 1970-2000

Table 3: Additional Cluster Characteristics

	Clusters				
Variable	1	2	3	4	5
Cross border illegally?	39.21	67.25	41.51	67.30	69.34
Trip duration (months)	240.67	148.81	148.38	112.09	137.33
Number of subsequent trips	1.27	2.65	1.27	1.64	1.99
Apply for legalization in the U.S.?	68.24	69.56	60.40	50.46	67.46

4.5 The Relationship of Clusters to Migrant Strategies and Outcomes

Up to now, we have interpreted the nature of clusters in terms of mean differences in the variables on which the clustering was based. If our interpretation of the migrants clusters is valid, then, we should observe systematic differences among clusters on other variables. Table 3 presents relevant data.

One important variable of interest in our case is whether migrants cross the Mexican-U.S. border legally. Theories of migration do not make any predictions on migrants' border-crossing strategies. Studies have provided empirical answers to this question (Massey & Espinosa, 1997), yet those findings have not been incorporated into major theories of migration. Analyses of variance were conducted to assess cluster differences in border crossing. We find that clusters differ significantly in their likelihood of crossing the border without documents ($F(4, 15065) = 283.81$, $p < 0.000$). Network migrants and risk-diversifying migrants are less likely to migrate illegally (only 40% do) compared with other clusters (at least 67% do).

Clusters also differ significantly in length of stay in the U.S. ($F(4, 15065) = 283.81, p < 0.000$) and the total number of subsequent trips ($F(4, 14559) = 340.86$, $p < 0.000$). Risk-diversifying migrants (cluster 1) are most likely to stay long, while income-maximizing (cluster 2) migrants are most likely to do subsequent trips. The difference among clusters is also significant in whether migrants apply for legalization in the U.S.($F(4, 9961) = 54.76, p < 0.000$). Compared with the other groups, pull migrants (cluster 4) are less likely to apply for a green card or citizenship.

5 Conclusions

These preliminary findings suggest that there are at least five distinct groups of migrants from Mexico to the United States. The results show that there is a clear time pattern to the relative dominance of each migrant type, which overlaps with macroeconomic and political trends in Mexico and the United States. Moreover, results suggest that migrant types may influence several other migration strategies and outcomes (legal or illegal entry, subsequent trips, length of stay). Despite these intriguing findings, several improvements to the ideas and analyses presented here are in order.

For example, future iterations of the paper will use hierarchical clustering techniques in addition to simple partitioning applied so far. More formal stopping rules, as an alternative to the ad-hoc measure defined in the paper, will be utilized to determine the number of clusters in the data. To achieve these goals, the analyses will be carried over from Stata to Matlab platform. (Stata has a strict matrix size restriction, which prevents one from applying the hierarchical clustering techniques to a large data set, as the one used in this paper. Furthermore, while there are numerous stopping rules suggested in the literature, Stata has only implemented two stopping rules. Its programming capabilities to introduce additional stopping rules are also limited. Matlab imposes no restrictions on data or matrix size, and provides a more flexible programming environment to implement additional clustering techniques and stopping rules. The code written in Matlab for clustering will be made available online for other users.)

This version of the paper includes only first migration moves of individuals. There are recently suggested clustering techniques that deal with time-series observations, which may allow us to look at all migration moves by individuals in the future iterations of the paper.

References

Bailey, K. (1975). Cluster analysis. *Sociological Methodology*, *6*, 59–128.

Everitt, B. S., Landau, S., & Leese, M. (2001). Arnold.

Fleishman, J. (1986). Types of political attitude structure: Results of a cluster analysis. *The Public Opinion Quarterly*, *50*(3), 371–386.

Massey, D. S., Arango, J., Hugo, G., Kouaouci, A., Pellegrino, A., & Taylor, J. E. (1993). Theories of international migration: A review and appraisal. *Population and Development Review*, *19*(3), 431–466.

Massey, D. S., Arango, J., Hugo, G., Kouaouci, A., Pellegrino, A., & Taylor, J. E. (1994). An evaluation of international migration theory: The north american case. *Population and Development Review*, *20*(4), 699–751.

Massey, D. S., & Espinosa, K. (1997). What's driving mexico-u.s. migration? a theoretical, empirical, and policy analysis. *American Journal of Sociology*, *102*(4), 939–999.

Miller, D., & Friesen, P. H. (1977). Strategy-making in context: Ten empirical archetypes. *The Journal of Management Studies*, *14*(3), 253–280.

Piore, M. J. (1979). Cambridge University Press, Cambridge.

Stark, O., Taylor, J. E., & Yitzhaki, S. (1986). Remittances and inequality. *The Economic Journal*, *96*, 722–740.

Todaro, M. P. (1969). A model of labor migration and urban unemployment in less-developed countries. *American Economic Review*, *59*, 138–148.