Agents and Decision Trees from Microdata

John B. Nelson, William G. Kennedy
George Mason University
4400 University Drive
Fairfax, VA 22030
JBN@pathdependent.com, WKennedy@gmu.edu

Ariel M. Greenberg
Johns Hopkins University / Applied Physics Laboratory
11100 Johns Hopkins Road
Laurel, MD 20723
Ariel.Greenberg@jhuapl.edu

Keywords:
Microdata, Probabilistic Graph Models (PGMs), Decision Trees, Agent-Based Modeling (ABM)

ABSTRACT: *This paper discusses the development of a model of the household migration behavior of a nation's population. From information synthesized from across available microdata sources which are each temporally, spatially, or topically inconsistent in coverage, we learned decision trees and instantiated agents in an agent-based model. The generative results of the whole-country simulation of this ABM mimicked the observed macro-level findings, engendering confidence in this method to develop agents and decision trees from microdata.*

1. Introduction

The Air Force Research Laboratory's National Operational Environment Model (NOEM) is an ambitious project to rapidly populate models of any arbitrary country from a wide variety of open source data, and enable *in silico* experimentation on these models. It integrates several model types. Most modules were first implemented as system dynamic models. A radicalization model, part of NOEM's behavior module, was implemented as an agent-based model adapted from (Epstein, Steinbruner, & Parker, 2002) and is still expanding to accommodate migration and crime related behaviors. In this paper, we summarize our effort over the last year in developing a data-driven agent-based model (ABM) of migration in the Republic of Colombia providing future behavior module capabilities to advance NOEM's migration and crime modules.

2. Data Sources

Broadly, we seek two classes of data to feed analysis and modeling: microdata and event data. This fine resolution is necessary if we assume heterogeneous decision making, a hallmark of agent-based modeling. Aggregate statistics are insufficient. We need to have realistic household socio-demographic variables and resource endowments. Here we enumerate those data sources ultimately employed.

1. IPUMS–International (Integrated Public Use Microdata Series, International) is a clearinghouse for microdata samples, which are anonymized but statistically valid samples from census data. For Colombia, the census source is DANE (Departamento Administrativo Nacional de Estadística). This is a very high resolution demographic and socio-economic sample. Roughly 1:10 households, geographically covering all of Colombia, are represented.

2. The Barometer series, Latinobarometer (LB) for Colombia, provides results from cross-sectional surveys that gauge public opinion on topics political, economic, and security-related.

3. Data Synthesis

A perennial challenge to performing social science research on collections from the developing world is data acquisition. Researchers interested in conducting country-specific modeling or statistical studies routinely encounter a dearth of particular microdata – survey or census data consisting of individual records from persons or households within a desired country, time frame, and topic set. On the rare occasion when such a set is available, it is usually from a small-N study.

This document is a product of work sponsored by the Air Force Research Laboratory (AFRL). Work is executed under the provisions of NAVSEA Contract # N00024-13-D-6400, Task Order # 169, Task ID # MC204. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressly or implied, of AFMC, AFRL, or the United States Government.

Microdata clearing houses such as the International Household Survey Network (IHSN), nationally-hosted microdata repositories like the Colombia National Statistical Office's DANE/ANDA, and topic-oriented regional surveys like the Barometer series (Asia, Afro, Latino, and Arab Barometer) provide microdata sets of independently conducted collections, each set distinct in time, topic, or geographic granularity (if not distinct geographic coverage).

Integrated sets in the spirit of the University of Minnesota's Integrated Public Use Microdata Series (IPUMS), like the NSF's Terra Populus and the NIH's Integrated Demographic and Health Series, harmonize within topics by identifying the fuzzy intersection of survey questions across time and space and grouping related questions into variables ranging from detailed to general.

Dataset fragmentation confounds research into interaction between topics covered by separately-sourced data sets, and stymies research into predisposing and precipitating factors that occur prior to data collection of the consequent behavior under study. For example, an exploration into the relationship between the topics of criminal victimization and migration behavior requires spanning barometer and census sets (topic synthesis), and a study into resource-driven migration requires backward imputation in time of the resources available to respondents during the period of their departure (time shifting).

Here we describe a methodology to synthetically unite microdata disparate in time and topic for use in statistical studies, and to seed realistic agents for agent-based models of complex social systems on which to run simulation experiments.

Such a methodology enables the production of a unified data assimilation pipeline (a synthetic repository of repositories) that compiles statistically reasonable synthesized microdata on demand. The keystone task in this vein is verification and validation (V&V) of synthesized data. Toward this end, we propose to first perform a survey of small-N studies and characterize each study by time, space, and topic coverage. This backbone of natively unified sets scopes which disparate large-N microdata sets may be synthetically unified. The sets synthesized from the identified large-N microdata are then necessarily amenable to be verified and validated at the record, aggregate, and results levels against the more inclusive small-N studies.

Time-shifting and topic synthesis operate on surveys collected within a particular country. Whereas one country's data is not reliably portable even to similar nearby counties, sourcing data from uncovered countries of interest cannot be overcome synthetically.

We introduce the novel idea that computational social science models can also serve as synthesis artifacts. Models require synthesized data; but they also can produce novel syntheses. Specifically, we propose to use probabilistic graph models to draw from diverse datasets and generate synthetic microdata samples, instantiated in agent populations, which encapsulate nonlinear relationships inferred from the source data. The basic process we used is shown in Figure 1.

4. Modeling

In modeling migration, there are two essential questions: what causes a household to migrate, and what criteria do they use to select their destination? We assume that, in making both of these decisions, households use information about their actual and relative circumstances (Hear, 2012). Unfortunately, the data on actual and relative circumstances prior to migration do not exist. To overcome this problem, we use Probabilistic Graphical Models as a means of estimating prior and comparative circumstances. Then, we construct decision trees that incorporate these estimates.

4.1 Probabilistic Graph Model

The rise of cheap and abundant computational power has made the use of probabilistic graphical models (PGMs) feasible. PGMs – and, in particular, Bayesian Networks (BNs) – are attractive in general for three reasons (Koller & Freeman 2009).

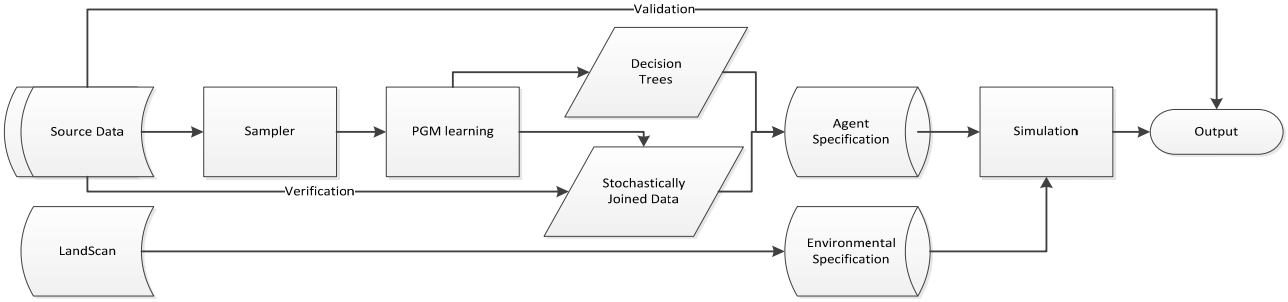

Figure 1 - Pipeline Process from Microdata through Simulation and Validation

First, PGMs represent statistical relationships concisely. Specifying a full joint-probability table for many variables quickly becomes intractable owing to combinatorial explosion, even ignoring the issue of statistical significance. By inferring conditionally-independent relationships, PGMs factor the event space into a manageable subspace. At the same time, PGMs are transparent (in contrast to opaque models, like artificial neural networks). Consequently, an expert can interrogate proposed graphical structures for face validity; suggest alterations; and, importantly, impose constraints implied by established theory.

Second, the graphical structure allows for fast inference and sampling, both in general and conditional upon some known evidence. Sampling conditionally upon known evidence (i.e. sampling over the posterior distribution) is particularly useful as a means of joining disparate models.

Third, in the case of BNs, there is an existing and growing collection of open-source and proprietary tools for learning both the structure and parameters of the models. This learning may be purely data-driven, deriving the most likely models from extant data; or, it may be inferred while incorporating expert opinion.

These attributes are especially attractive for social science research, where control and prediction are subordinate to satisfactory explanation. Fidelity of proposed relationships matters, and a tool that is transparent, fast, and amenable to theory building and hypothesis testing is very useful. Recognizing the promise of PGMs for social science research, BNs are experimentally employed for this project in three distinct but related ways.

While IPUMS provides only cross-sectional data, certain survey years have variables indicating the respondent's previous residence (sub-national regions or both sub-national regions and urban areas). This variable grants a toehold for imputing the most plausible explanation (MPE) for the respondent's previous characteristics, casting it from cross-sectional to quasi-longitudinal. First, a collection of BNs is learned for each urban area represented in IPUMS, limited only to non-migrant respondents. Then, these responses are partially-translated backwards through time, adjusting deterministic variables when possible (e.g. subtracting five years from their age) and marking other variables as unknown. The known variables – those that are deterministic or immutable (e.g. sex) – are used as evidence when querying the learned, urban-level BNs. This procedure effectively asks: assuming the respondent is not exceptional and that municipalities change slowly (over many years), what values could we expect the missing values to take for migrants, in the statistical sense? Once imputed, a new set of BNs can be learned which infer the statistical structure of observed and imputed variables at the previous and present sites.

This same general procedure – sampling from an existing BN conditional upon some fixed evidence – can be used to join disparate datasets. Generally, when two different datasets share a set of common variables such as the respondent's socioeconomic status, municipality, gender, age, family size, and number of children, a merged record can be synthesized. A record (sampled from the BN or drawn from survey data) from one dataset is taken as prior information; then, the BN inferred for the second data source is sampled, conditional upon the values of variables in the union of the two sets – the evidence. The result is a statistically plausible merger of two disparate data sources.

In addition to their use in exploratory data generation, the collection of PGMs is used to seed the agent-based model with survey data. The alternative – assuming a non-existent source of longitudinal and cross sectional survey data – would be to initialize the model statically. By using the PGMs, agent initialization is empirically constrained, yet stochastically determined. And, it also allows for redistribution, something typically prohibited with raw data.

IPUMS	SEX	AGE	Has AUTO	HOME OWNRSHP	HAS PHONE
LB:2000	S1	S2	S10G	S10C	S10F
LB:2001	S1	S2	S10G	S10C	S10F
LB:2002	S1	S2	S10G	S10C	S10F
LB:2003	S1	S2	S10G	S10C	S10F
LB:2004	S1	S2	S10G	S10C	S10F
LB:2005	S6	S7	S15H	S15C	S15F

Table 1. Mapping Join Variables to Source Variables

While survey data sets with a large sample size are exceptionally rare, small-N surveys that ask a broad set of theoretically interesting and useful questions are less rarely available. Since the synthesized data can be sampled easily and dynamically, these small surveys can be used for verification and validation of the amalgamated models. Recognizing that the statistical significance of small-N surveys is limited – especially when asking questions about the joint distribution – it is still useful to ask: how well does the synthesized model comport with the small set of observations? If they are statistically similar, this provides evidence of fidelity with the real world (see section on Verification & Validation). Absent demonstrable verisimilitude, it allows the researcher to ask what relationships might have been missing, which assists theory building.

4.2 Synthetic Respondents

In answering the modeling questions, the first task becomes: what did the respondents' lives look like prior to migration? IPUMS is the largest extant source of harmonized microdata for Colombia. Yet, as mentioned previously, it is purely cross-sectional. However, the existence of columns giving the prior residence for individual respondents provides the aforementioned toehold for generating estimates of prior conditions. From this column, and using PGMs, the desired responses can be synthesized.

First, the set of respondent household heads who did not migrate in the five years prior to survey is collected and partitioned by municipality. As non-migrants, there is no geographic shift for this group. There is only a shift in time, affecting their age and possibly number of children. For example, if it was reported that they had a child less than five years of age at the time of survey, their number of children would need to be decremented when assessing their situation five years prior. It is certainly possible that their resources would have changed during this period. However, for the sake of imputation, it is assumed that their resources were static. Roughly, this translates into the assumption that for non-migrants, little changes in five years; but, since we know they were stationary in geographic space, treating their resources as fixed gives us a means of estimating resource levels by municipality for non-migrants. A set of PGMs – one for each municipality – is trained on the data in this partition.

Then, the set of respondents who *did* migrate in the five years prior to survey is collected. This group is then shifted back five years in time and placed in the municipality of their previous residence (as previously discussed). However, we cannot assume that their prior resources are static, as in the case of non-migrants. They have shifted geographically, and different municipalities have very different resource profiles. Instead, for each migrant, the geographically specific, non-migrant PGM is queried with six variables as the evidence keys. The first three are:

1. *AGE_GROUP* is a mapping from their recorded age to a categorical value: [0, 20) => 0, [20, 30) => 1, [30, 40) => 2, [40, 50) => 3, [50, X) => 4. This mapping was necessary to reduce the parameter space for the PGM over age. Age value is important, but if a query were conditioned by a value bounded by 0 and 100 in integer space, there would be too few observations for some events, even when backed by larger data sources. This variable changes, but by a deterministic rule with respect to time.
2. *IS_MALE* is a predicate with *true* signifying the respondent was male. This variable is predominantly constant with respect to time.
3. *HAS_AUTO* is a predicate with *true* signifying the respondent had at least one automobile. This variable was included as a proxy for resource wealth, but it is a portable asset so it is reasonable to assume it follows migratory respondents.

Querying the departure PGM with the evidence variables provided by the extant microdata yields a statistically probable, synthetic observation. That is, the evidence parameters are taken as fixed, and the PGM returns a set of observations for the non-evidence fields that is justifiable given the structure learned on the empirical data. Or, in the context of migration, it yields a set of resources and other variables that were expected for the migrant in the departure site, given their fixed parameters.

After the migrants were mapped to a set of synthetic records at the departure site, a new series of PGM were generated. This series learned on the concatenation of the synthetic migrant records and the observed non-migrant records in the departure municipality. This PGM is subsequently used to seed the agent-based simulation.

In addition to back-imputed PGMs developed from the IPUMS microdata, a series of PGMs were constructed for the Latinobarometer (LB) data. These data have a much smaller sample size, with roughly 600 household head observations per year between 2000 and 2005, inclusive. A PGM was constructed by learning the structure and parameters for each year, separately. These PGMs are used to update the respondent profiles on an annual basis, at a national level.

Ideally, a national-level PGM could be used to express prior probabilities, joined in a Bayesian fashion to produce an estimate for each municipality given the observations. This is necessary, because the geographic area covered -- 153 unique municipalities – disperses the sample size significantly. With a prior model (an extension discussed in the Prior Models recommendation), the PGM could be specific to each municipality, as in the case of IPUMS. However, since the geographic coverage is neither exhaustive for Colombia, nor as extensive as IPUMS, additional imputation would be required.

To provide this imputation, the data could be geographically filled in using the IPUMS data to find similarities between municipalities. First, the set of municipalities would be partitioned into two sets: one set for those with overlapping coverage between IPUMS and LB (the candidate set) and another with only IPUMS. To fill the latter set, the most similar municipality from the overlap set would be identified; then, the PGM for this municipality would be transplanted.

Selecting the metric for similarity requires more empirical work, yet there is an obvious candidate. First, the candidate set would be filtered so that only those with similar levels of urbanism are retained. Then, from this reduced set, the municipality with the most similar proportions of migration would be selected. The assumption here is that municipalities experiencing similar migratory events were subject to similar processes. This is a strong assumption, but in absence of additional observations, it is a required step.

Although the national-level PGMs were less geographically specific, they should be capable of capturing patterns of victimization, and the relationships between the provided variables. Using the national-level PGMs, Latinobarometer was stochastically joined with the IPUMS to generate rich, synthetic responses. To join the two PGMs, a sample is first drawn from the IPUMS PGM in a specific municipality. Then, five of these variables are used as evidence parameters when querying the Latinobarometer PGM in a specific year: IS_MALE, AGE_GROUP, HAS_AUTO, OWNS_DWELLING, and HAS_PHONE. (See Table 1 for a mapping of LB source variables to the model's variables, by year.) OWNS_DWELLING is a predicate indicating home ownership by the household head (identified by SIDP as a predictive factor for victimization). HAS_PHONE indicates that the respondent claimed to own a phone, a critical variable for establishing respondent communications capability. Although the Latinobarometer PGM was national level only, the synthetic joined record benefits from the localization provided by the IPUMS PGM. Home ownership and telecommunication networks vary greatly by municipality, which is captured by the IPUMS PGM.

The result of this procedure is a set of tools capable of generating empirically plausible sets of synthetic responses that are geographically and temporally localized. These data are otherwise unavailable, with no surveys conducted with similar depth and breadth.

4.3 From Observed Patterns to Decision Trees

The set of IPUMS-based PGMs constructed as a means of temporally and geographically seeding and updating agents in the agent-based model were also used to interrogate the logic behind respondent migratory decisions outside the ABM. The set of non-migrant, raw IPUMS responses were taken as given. Additionally, a set was constructed for each respondent in IPUMS that did migrate. These migrants were moved back in time and space, and joined with resource profiles that were statistically probable for them. Then, the two sets were concatenated, representing a statistically probable survey for all respondents at their departure sites. This concatenation was then used to analyze two broadly identifiable decisions: the decision to leave in general, and the selection of a destination site.

The first question – succinctly, who migrates? – proved to be more difficult to answer. Seemingly, there are many idiosyncrasies in choosing to leave your residence, excepting expulsion. However, some general patterns became apparent. Curiously, the population of migrants appear better educated, wealthier, and more connected than the population of non-migrants, even after accounting for urbanism.

However, non-migrants are not necessarily the least wealthy group. When looking at non-migrants against all migrant types – migration due to violence, work needs, family needs, and health needs, and college – we find a remarkable similarity between migrants fleeing violence and non-migrants. There are several candidate explanations for this observation. First, the inference engine struggles with violent migration, because it is the statistically smallest sample of the group. Second, the inference engine struggles with violent migration, because the stochastic join variables failed to capture some essential discriminating factor. Finally, *migrants who are moving for a job, college, health, and family are the socioeconomic elite.* Everyone else is part of the lower to lower-middle class (henceforth, the disadvantaged), which should be the largest group with respect to socioeconomic inequality. The disadvantaged are more likely to be victims; but, violence and victimization has a strong spatial element. Therefore, the socioeconomic variables may not weigh heavily when looking at the population from 10,000 feet. It just means they are predisposed to being victims of violence, if violence arrives. Retaining only the set of migrants and partitioning by imputed cause of migration, more patterns become apparent.

In general, the decision to migrate appears to be stochastic. Observed characteristics predisposed some groups to migration, but an agent-centric decision structure remained elusive. However, the second question – of which destination is chosen – yielded more concrete answers. To set up the problem, each known migrant was backwards imputed to his or her origin site. Then, a set of seven candidate destination sites was collected for each respondent. These candidate sites were composed of the two largest nearby municipalities; the two smallest nearby municipalities; two random municipalities; and, the site actually chosen, as given by IPUMS. Here, *nearby* means within a few hops, where a hop of 1 spans contiguous municipalities. If the chosen site was included in the set of six candidate sites, an additional random site was included. The two largest sites were included given extant literature demonstrating the robustness of gravity models (Peeters, 2012) with population as an attractor. The smallest sites were

included as contrasts. And, the random locations were used to ensure sufficient variance.

For each destination site, the expected value for each variable was generated by sampling over the destination-site PGM. This expected value could be thought of as the agent's expectation if they were to move to a considered site. That is, given their age, gender, resources as proxies by owning a car, literacy, education, and household structure, what resources (including employment) could they expect to have at the candidate destination? The expectation for each variable was transformed into a rank ordering between the individual respondent's candidate expectations. For example, the municipality that represented the least likely probability of having a job for them would be ranked 1 and the most likely would be ranked 7. These ranks were computed with respect to each variable independently.

The destination site reported in the raw data would be flagged as their choice amongst the set of candidates. This synthesized choice data was then fed to the CART decision tree learning algorithm (Breiman, et al. 1984). Classification algorithms occasionally produce bad results when there are unbalanced classes; that is, if there is one class that dominates the other in terms of frequency of occurrence (Japkowicz, 2000). Here, there are six "NOT-CHOSEN" and one "CHOSEN" for every respondent, thus qualifying as unbalanced. To alleviate this problem, the "CHOSEN" record was paired with one random record from the respondents six other "NOT-CHOSEN" records. This balanced the classes for presentation to CART. However, the underlying ranks were still computed with respect to the full candidate set, so the rank comparison structure is retained.

The resulting decision tree identified patterns in migrant destination site selection, contingent upon the reason for their migration. Following the specification of the CART algorithm, the tree is constructed from root to leaf in a manner to maximize information gain at every step. Effectively, CART is a tree search algorithm; the decision nodes are selected if they provide the best reduction in entropy, from amongst the set of candidate tests. Or, said more simply, it picks the condition that best separates the classes, then moves on to the next level of the tree. This procedure explains why variables expected to be explanatory are sometimes absent. If there is co-linearity between variables, then the classifier will find little gains to be made once the first variable had been selected. For example, if HAS_ELECTRICTY was selected as a node, then HAS_PHONE may subsequently provide little additional explanatory power, as they are highly correlated.

For all migrant types, the population of the destination was identified as the first node in the decision tree. Effectively, this means that the destination population best partitioned the space of chosen and not chosen candidate sites. This is in agreement with the extant literature on gravity models in migration. The decision trees conditioned by migration type follow.

Figure 2 depicts the truncated decision tree for households reporting migration due to "Family" considerations. The condition is portrayed with a square box. If the condition is true, the left path is followed; if it is false, the right path is followed. The leaf nodes (ovals) show the number of IPUMS households falling into this category (n) and the odds of selecting this destination site.

To assist in theorizing, another set of decision trees was learned on the Latinobarometer data, classifying which respondents experienced victimization. Note, these trees did not require PGM generation or backwards imputation. They were learned on the raw survey data. When the CART algorithm was endowed with all the available variables, corruption was the best classifying node. This is reasonable – people who have experienced corruption seem more likely to have experienced violence. Ignoring corruption, the size of the town, access to drinking water, gender, age, and whether or not the respondent read the newspaper (information provenance), were the algorithmically selected nodes (For space considerations, this set of trees are not depicted). Perhaps, whether or not the respondent read the newspaper affected the perception of crime rates.

Another decision tree was limited to perceptual variables. Again, the size of the town was the most important factor, and reading a newspaper remained important. However, employment, perceptions of future economic conditions, and their perception of whether their children would be would live better

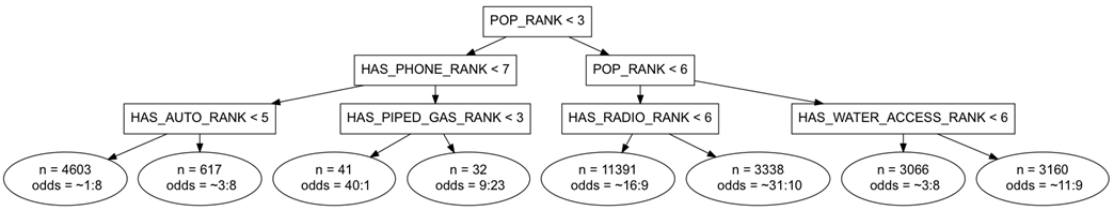

Figure 2 - Migration Decision Tree

comprise the next layer of nodes. Seemingly, violence and the perception of violence affect prospective evaluations more than they do retrospective ones, even though the event of violence should be attached to the retrospective evaluations, in that they occurred already.

5. Simulation

In the previous section, we described our approaches to producing synthetic respondents, and their decision trees. The products of these approaches furnish the properties and cognition (respectively) of the agents and behaviors to be simulated.

5.1 Homeland Model

HomeLand, our agent-based simulation of migration and victimization in Colombia, is described in detail in Kennedy, et al. (2014) using the ODD+D approach (Muller et al., 2013). It was built in Java using the MASON agent-based modeling environment (Luke, et al. 2005). The nearly 5 million agents represent individual households of the entire population of Colombia and are located in 1km^2 parcels based on the population density. Each household's location is within a municipality within department (the equivalent of US counties and states) and initialized using population data. To support social communication among our agents, we extended MASON by building a message exchange layer (MEL). Each household considers whether to migrate and where they might migrate each step representing a year. The results of that decision process may or may not be communicated as a recommendation to the household's social network made up of their immediate neighbors, more distant neighbors, and a few households selected at random from the entire country.

5.2 Experiments & Results

Our experimental design is 2 × 2, given in Table 2 with conditions marked with short names. We generate migration behavior under conditions in which household decisions are, or are not, informed by communications from other households, and decide where to move with or without using perceptual data. The social condition tallies recommendations from the household's social network on whether or not to migrate. The perceptual condition uses annual Latinobarometer survey data to evaluate possible destinations and that data includes perceptions of relevant trends, such as whether households that move there think the economic conditions are better. Simulation output is reported in the NOEM input/output format, the schemas for which are given in the tables below (values available on request). As is, constituent values are produced at the household level, but household members could be retrieved reconstruct the full population.

	Without communications	With comms
Excluding Perceptual Variables	Base	Social
Including Perceptual Variables	Perceptual	Social + Perceptual

Table 2 - Experimental Conditions

5.3 Verification & Validation

Verification and validation is conducted here in the spirit of model clamping described above. Input to the simulation is verified by comparing synthesized respondent data to the source respondent data from multiple original sources. We perform this comparison of mortality and employment (from IPUMS) at the municipality level for the year 2005, which each set and the simulation have in common. Comparison of victimization between LB and the synthesized set is challenged by few observations by municipality in LB, and LB's low statistical representativeness of the population's demographics.

	Synthesized Set
IPUMS (mortality)	0.83
IPUMS (employment)	0.93

Table 3 - Data Set Agreement (Regression, R^2)

Output from the simulation (population counts, migration) over the course of the simulated period, 2000-2005, is validated at the department level against the 2005 endpoint captured by IPUMS. The final population counts by department over the five year simulated period for each of the four conditions are in reasonable agreement, excepting the occasional outlier as shown in Figure 3.

6. Discussion and Conclusions

The final population counts by department over the five year simulated period for each of the four conditions are in reasonable agreement, excepting the occasional outlier. Where the *baseline* and *social* conditions agree closely with IPUMS counts, the conditions *social + perceptual*, and *perceptual* tend to underestimate. Where the *baseline* and *social* conditions overestimate or underestimate IPUMS counts, the conditions *social + perceptual*, and *perceptual* tend to be closer to the IPUMS counts.

This pattern of compensation between model conditions that take perception into account with those that do not suggests the possibility that the particular cognitive scheme for an agent to employ may itself be selected according to personal or environmental characteristics. For example, if conditions are particularly fearful, stressful, or dreadful, perception may impact decision-making more than under conditions that are bad, but not so extreme

In this paper, we documented the simulated reproduction of migration behavior as collected by the Colombian census. Much of our effort was in the integrating data from small surveys, modeling and analyzing that data, and developing representations of the decision-making of the households' internal migration. This was then used in a simulation and confirmed to match the available data.

While this project was primarily an effort to conduct advanced development for NOEM's migration and crime modules, two methodological capabilities applicable to both modules as well as the behavior module emerged along the way: microdata set synthesis to seed agents, and the production of decision trees from these synthesized sets to endow the agents with empirically-grounded models of decision making. We suggest these methods are useful in developing data-driven social simulation.

7. References

Breiman, Leo; Friedman, J. H.; Olshen, R. A.; Stone, C. J. (1984). *Classification and regression trees*. Monterey, CA: Wadsworth & Brooks/Cole Advanced Books & Software.

Epstein, L. M., Steinbruner, J. D., & Parker, M. T. (2002). Modeling Civil Violence: An Agent-Based Computational Approach. *Proceedings of the National Academy of Sciences*, *99*(suppl 3), 7243–7250.

Hear, N. Van, Bakewell, O., & Long, K. (2012). Drivers of Migration. Migrating out of Poverty Research Programme Consortium, (March).

Ibáñez, A. M., & Vélez, C. E. (2008). Civil Conflict and Forced Migration: The Micro Determinants and Welfare Losses of Displacement in Colombia. *World Development*, *36*(4), 659–676.

Japkowicz, N. (2000), The Class Imbalance Problem: Significance and Strategies, in Proceedings of the 2000 International Conference on Artificial Intelligence (IC-AI'2000), pp. 111-117.

Kennedy, W.G., Nelson, J.B., and Greenberg, A.M. (2014) HomeLand: Overview, Design Concepts, and Details (ODD+D), Rev 10, 20 Dec. 2014.

Luke S., Cioffi-Revilla, C., Sullivan, K., Catalin Balan, G. (2005) MASON: A Multiagent Simulation Environment. Simulation 81(7). pp. 517—527.

Müller, B., Angermüller, F., Drees, R., & Dressler, G. (2013) Describing human decisions in agent-based social- ecological models - ODD+D an extension of the ODD protocol, 1–39.

Peeters, L. (2012). Gravity and Spatial Structure: the Case of Interstate Migration in Mexico*. Journal of Regional Science, 52(5), 819–856. doi:10.1111/j.1467-9787.2012.00770.x

Minnesota Population Center. Integrated Public Use Microdata Series, International: Version 6.3 [Machine-readable database]. Minneapolis: University of Minnesota, 2014.

The authors wish to acknowledge the statistical office that provided the underlying data making this research possible: National Administrative Department of Statistics, Colombia.

Latinobarometer Corporation, "Latinobarometer 2000-2005", http://hdl.handle.net/1902.1/10611 V1 [Version]

Author Biographies

JOHN B. NELSON is a Ph.D. student at George Mason University. He is interested in computationally modeling belief systems, especially with respect to American voting.

WILLIAM G. KENNEDY, Ph.D., Captain (USN, Ret.) is a research assistant professor with the Department of Computational Social Science, part of the Krasnow Institute for Advanced Study, George Mason University.

ARIEL GREENBERG is a Project Manager in the Intelligent Systems Center of JHU/APL. His research spans the domains of computational biology and social science, including topics in psychophysiology and behavioral modeling. Ariel received degrees in Biology and in Electrical Engineering from University of Maryland, College Park.

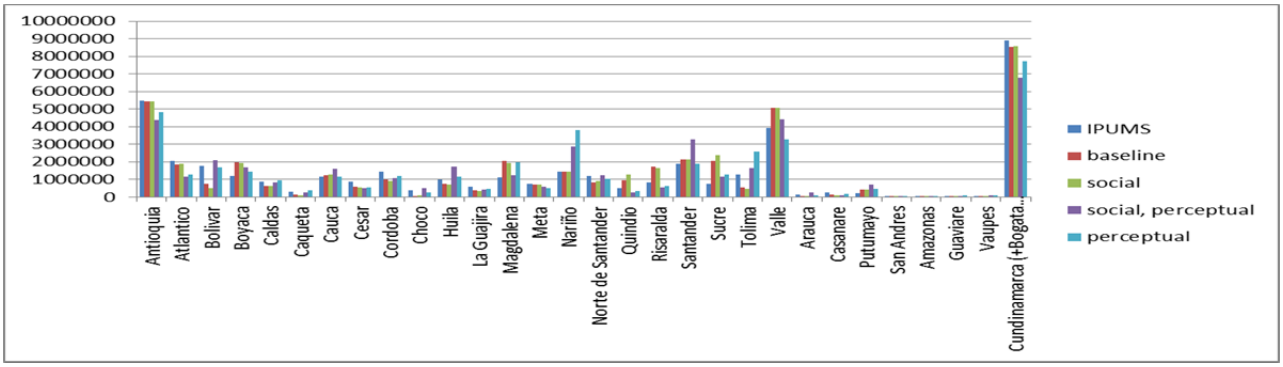

Figure 3. Validation Comparing Simulation Results to IPUMS Survey at the Department Level