Construction of Synthetic Populations with Key Attributes: Simulation Set-up while Accommodating Multiple Approaches within a Flexible Simulation Platform

Steven J. Fernandez, Amy N. Rose, Edward A. Bright, Justin M. Beaver, Christopher T. Symons, and Olufemi A. Omitaomu, Cathy Jiao

Oak Ridge National Laboratory

Abstract—In this paper, we describe our concept for overcoming the data barriers of building credible synthetic populations to assist the transformation between social theories and mathematical models. We specifically developed a 31-million-agent model of Afghanistan’s population to demonstrate the ability to computationally control and analytically manipulate a system with the large number of agents (i.e., $10^8$) necessary to model regions at the individual level using the LandScan Global population database. Afghanistan was selected for this case study because gathering data for Afghanistan was thought to be especially challenging. The LandScan Global population database is used by a majority of key U.S. and foreign agencies as their database system for worldwide geospatial distribution of populations. Assigning attributes to disaggregated population was achieved by fusing appropriate indicator databases using two forms of aggregation techniques—geographical and categorical. A new approach of matching attributes to theoretical constructs was illustrated. The other data sources used include data on military and peacekeeper forces’ loyalties, readiness, and deployment collected through a combination of UN and classified force projections; economic data collected at the national level and disaggregated using data fusion techniques; data on social attitudes, beliefs, and social cleavages through anthropological studies, worldwide polling, and classified sources; and data on infrastructure and information systems and networks.

Index Terms—agent-based models, social modeling, high-resolution data, population database, data fusion.

I. INTRODUCTION

One of the challenges in agent-based social modeling is identifying and characterizing global data sources as well as the granularity needed to synthesize and model populations with sufficient detail to support social-simulation environments. Data describing worldwide demographic, economic, and social attributes of the world population have previously been deemed inadequate for these purposes. However, geospatial sciences have advanced a number of techniques for merging disparate databases to achieve new levels of granularity for social simulators [1, 2].

The objective of this project was to identify and characterize global data sources and the granularity needed to synthesize and model populations at the level required to support the simulation environment. The metrics for this project are to:

- Demonstrate the direct use of observed or simulated data to seed the behavioral dynamics governing heterogeneous agents representing the entire Human, Social, Behavioral, Cultural (HSBC) domain at each scale of aggregation.
- Demonstrate the adequacy of available public domain or classified data sources that relate directly to individual human behavior as well as overall social or economic phenomena, with the objective of guiding the formulation of Complex Adaptive System (CAS)-based models capable of generating actionable predictive insights and modeling surprising emergent behavior.
- Demonstrate the adequacy of disparate, multiscale, and possibly dynamic data sources that can validate the causal patterns and the predictive insights extracted from CAS-based models for social or economic systems.
- Demonstrate the feasibility of using the observed and simulated data in a scalable fashion within a realistic modeling and simulation environment. Specifically, demonstrate the feasibility of enabling the online data-model feedback loop in an environment with more than 100 million agents, with complex behavior based on realistic social or economic theory, and with disparate data feeds coming in at real-time or near real-time.

In summary, this project seeks convincing evidence that the data and related tools exist to construct a synthetic population for the simulation tasks.

The remainder of this paper is organized as follows. In Section II, we discuss the definitions and descriptions of some of the data used. In Section III, we present global visual
As a first step, we assessed data quality and then illustrated the concepts by constructing a synthetic population. The criteria used to assess the data quality follow.

II. DATA DEFINITIONS AND DESCRIPTIONS

As a first step, we assessed data quality and then illustrated the concepts by constructing a synthetic population. The criteria used to assess the data quality follow.

A. Definition of Data-Quality Levels at the National Level

The criteria for assigning a data quality level are based on the ability to use the assembled data sources to build synthetic populations. Scores of 0, 1, 2, and 3 are assigned to variables describing geographical territories. This scale reflects the approximate difficulty in building synthetic populations that support a wide range of potential social theories.

Areas assigned the value 0 have no data available for building synthetic populations. Examples of such areas include Antarctica and the Australian Indian Territories. Areas assigned a value of 1 have below-average data quality. In general, these countries are characterized by infrequent or nonexistent census activities and include the central African countries; countries whose data submissions to the United Nations are under suspicion of gross manipulation, such as Cuba and some Former Soviet republics; and a few Central and South American countries. Conflict zones typically fall within this category because of the transient nature of their populations and the difficulty in collecting data in conflict zones. Exceptions include areas where U.S. interest in the political and social stability is high (e.g., North Korea, Iran, Iraq, and Myanmar). Areas assigned a value of 2 have average data quality. These countries routinely submit data to the United Nations, and the data is generally useable with only ordinary quality-assurance checking. Census surveys and data reporting may be less frequent than in North America or the European Union. This category includes most of the developing countries and relatively small participants in world trade flows. Areas assigned a value of 3 have better-than-average data quality. Census, military structure, economic, social, and infrastructure data are collected annually with known accuracy and precision. This category comprises North America and the European Union, members of the OECD, and major trading partners.

B. Definition of Independent Variables

The data required to build a synthetic population can be organized into six broad dimensions—political, military, economic, social, information, and infrastructure—that are important for understanding how populations react and respond to changes in their environment. Within each dimension, the project social theorists surveyed various models and cataloged the independent variables that would satisfy a large fraction of the modeling and surveying needs. Therefore, different data sources for independent variables were identified.

C. Assigning Data-Quality Levels at a Dimension Level

Within a dimension of the independent variables, the following criteria were used to assign the data-quality values.

Political. Almost all countries meet the criteria for data quality level 1, which is identifiable national legislative, executive, and judicial bodies, roles, and representation. Data at this level are of equal or better quality than data in the CIA Fact Book. Data quality level 2 is reached if data exist to interpolate or derive political attitudes and beliefs. Regular polling, census, or anthropological studies that define a time series of loyalties may provide this data. Likewise, information on subpopulations that is derived from data on social cleavages (as determined from social variables) may enable sub-national detailing of political ideology, freedom of action, and resources. Data quality level 3 requires that data be available to describe or interpolate the regulation and actions of governmental and nongovernmental organizations as well as their motivation, intent, and objectives. Most close U.S. allies meet the criteria for data quality level 3.

Military. The military dimension of a synthetic population can be separated into two components: (1) the make-up, reliability, and capability of the country’s armed forces and (2) attitudes and beliefs about a country’s security. Values of 0 or 1 are assigned to areas of active conflict for which little data exists on the make-up, structure, or force level of combatant forces, the acceptance of these forces by the population, or the popular attitudes and beliefs regarding these forces in their roles as peacekeepers, humanitarian support, or a positive influence on general well-being. A data quality level 2 is assigned to countries participating in UN peacekeeping activities or to countries where other data allows a social cleavage analysis of the armed forces. A data quality level 3 is assigned to countries that have entered into mutual security arrangements with the United States and to countries that have been topics of a classified National Intelligence Estimate to address these factors.

Economic. The data quality levels for the economic dimension are highly dependent upon the occupational distribution of the population and the granular distribution of economic activity within the country. Therefore, the assigned quality level for each country was no greater than the LandScan Global data quality level for that country. Members of the World Bank and countries contributing data to the International Monetary Fund development and statistics database typically have data that, when combined with other sources, is sufficient to address economic welfare and a leader’s ability to allocate resources. Exceptions are nations that have been the subject of classified National Intelligence Estimates that address these factors. In those cases, a data quality level 3 has been assigned.

Social. Data in the social areas can be considered a fusion of data from other domains. However, we have called out this dimension separately because many of the interesting social theories are highly dependent on the variables in this domain. The social dimension depends not only on this fused data and open sources of collected data, but also is heavily dependent
upon attitudes, beliefs, and cultural values that can be developed only by Human Terrain methodologies. As a proxy for these attitudes and belief mapping, we used Gallup World Poll data and the Gallup methodology to make an open assessment of the information available to populate the attitudes and beliefs of a synthetic population. A data quality level of 1 was assigned if the Gallup World Poll was not available within that country and supplemental data sources were not available for human terrain mapping; a level of 2 was assigned if the country had only Gallup World Poll coverage; and a level of 3 was assigned if the polling coverage was supplemented with continuing coverage from additional sources.

**Infrastructure.** The data quality levels for infrastructure were based primarily on engineering sufficiency of the data to allow the performing community to build models of the key infrastructure sectors. For most sectors, the necessary data exist within the U.S. Department of Defense (DoD) mission planning community. This all-source data is currently being supplemented by National Geospatial Intelligence Agency’s (NGA) open-source infrastructure resources. For example, NGA’s Palanterra, a Web-based interface designed to describe, assess, and depict physical features and geographically referenced activities, integrates disparate stores of spatial data to create a spatially and analytically enabled operational picture. The data quality level reflects a qualitative assessment of the data available through Palanterra.

**Information.** The criteria for data quality in the information dimension were based on a qualitative assessment of the ability to model the telecommunications infrastructure sector. For most of the sectors, the necessary data exists within the DoD mission planning community. The open-source baseline data are available in the World Bank’s Little Data Book on Information and Communication Technology 2007, and this is supplemented by resources like NGA’s Palanterra.

### III. Visualization of Data Quality by Dimension

Country-by-country data quality levels as assessed using the criteria in Section II are shown in the following maps (Figure 1) for each dimension.
IV. THE AFGHAN EXAMPLE

As an example, we constructed a synthetic population representative of Afghanistan’s 2006 population, with the ultimate objective of testing the ability of various leadership theories to answer strategic questions. One of these questions asks how to reduce the overall level of terror incidents. In this case, the leadership theories require input on loyalty, coercion, ideology, economic welfare, and security from violence for each agent of the population. To geospatially assign attributes for the agent’s support for the leadership, the leadership’s resources, the agent’s ideology, and the agent’s degree of satisfaction with the economic and security situations, databases and data sources were employed according to Figure 2.

The first step in building the synthetic population was geospatially locating the 31 million agents representing the Afghan population within grids approximately one kilometer on each side. The LandScan Global techniques for disaggregating a population produced the geographical population distribution illustrated in Figure 3.

Assigning attributes to this disaggregated population was achieved by fusing appropriate indicator databases. A diverse body of literature describes the major Afghan political and social cleavages that can be used to deduce ideology and perceptions of security against violence. We began by consulting district profiles collected through human-terrain-type efforts by the Afghan Information Management Service (AIMS) from which we can determine the ethnicity of each element in 206 of the 329 Afghan districts. For the remaining 123 districts, the ethnic breakdown was estimated by

Fig. 2. The flow of data source employment in construction of the Afghan synthetic population

Fig. 3. Geospatial distribution of agents representing the Afghan population is represented on one kilometer cells.
interpolating data on ethno-linguistic maps, such as in Figure 4 that delineates the distribution of ethnic groups. Ethnic data from adjacent districts was used for the interpolation when it was available. Otherwise, data from nearby areas sufficed. Although many granular maps exist, the National Geographic maps were used to demonstrate the value of common maps of limited granularity.

Once the ethnicity for each of the 31 million agents was assigned, then the ideology and attitudes toward the economic and security situation were assigned to each social atom.

The attitudes within the Gallup study group seem to have significant cleavages along ethnic (Pashtun, Tajik, and Other), location (north, south), and combat proximity (war zone vs. no-war zone) lines. A mapping for war zone location is reproduced in Figure 5. The indicator database for how economic welfare is perceived by the individual and is tied to behaviors in support of leaders in the Afghan case is highly dependent upon the social atom’s proximity to opium production (Figure 6). The elements above, taken together, were the building blocks of our synthetic population.

V. LESSONS LEARNED AND INSIGHTS

In the effort to demonstrate, or document the feasibility of, direct use of observed or simulated data to seed the behavioral dynamics governing heterogeneous agents representing the entire HSBC domain at each scale of aggregation, we have learned it is feasible. Although assigning non-observable beliefs and attitudes to individual social atoms remains beyond the state-of-the-art, modern demographic techniques such as those embodied within NGA efforts can build synthetic populations with a low number of clones. As a next step, we attempted to correlate the data sources with attributes required to test proposed social theories in the area of responsibility.

For an example we used a Nigerian area of responsibility. In this case, the simulation’s objective question reflected insights about disruption of oil flows from the Niger Delta. In the second step, developmental data sets were constructed suitable for constructing a synthetic population for the simulation system. The attributes of the synthetic population assigned to each agent of the simulation is dependent upon the scenario elements or actions that make up a proposed course of action (COA) and the independent variable of the social theory selected for testing within the simulation system.

This approach requires the data sets and repository to be constructed in three parts: 1) a data set of scenario elements representing plausible COAs built by analysts based on the simulation objective; 2) a data set of social theory elements, defined as memes (element of cultural ideas, symbols or practices), constructed through application of theories to the area of responsibility, then deconstructing the theories into a data set of memes suitable for correlation with the scenario elements using national language processing; 3) a data set of evidence representing real-time or archived records of events, analysis, and analytical reporting from and about the area of responsibility. These three sets, were correlated using Natural Language Processing (NLP) tools such as the Oak Ridge Piranha system.
The correlation step required an initial ontology of entities, relationships, actions and objects created by subject matter experts (SMEs) as a seed. In the absence of other information, this initial ontology is built from the scenario elements.

Scenario elements were built as simple text records to facilitate entity extraction and ontology construction. An equivalent data set of social theory memes were built from expert application of the essential elements of the following nine HSBC theories and was broken down for the theory extraction so that the knowledge extraction module may correlate the motivational elements to the scenario elements. The theories used for this pilot study are:

- Social Movement Theory
- Social Identity Theory
- Collective Action Theory
- Deviant Legitimization Theory
- Attribution Theory
- Social Cognition Theory
- Group Dynamics Theory
- Prospect Theory
- Conflict Theory

A subset of the above theories was placed within the context of area of study and broken into theoretical memes for natural language processing. At the conclusion of this task, a piranha-like cross mapping between the scenario elements and the motivational elements was ready to be compared to the raw evidentiary database.

The raw evidentiary database is resident at Oak Ridge National Laboratory and is composed of the following elements: characterized terrorist incidents structured according to the World Wide Incident Tracking System ontology and news reports describing political, security, economic, social, infrastructure, and information developments within the area of interest as well as unstructured messages providing data, information, and analysis in various stages of production and processing.

The correlation and cross-mapping of the evidentiary material was a pre-processing step to the likelihood inference generation. At this point, ingestion of the likelihood estimation into existing host tools to evaluate courses of action provides the hand-off point to the existing host analysis structure.

In this project, we demonstrated that scenario elements of courses of action could be rapidly correlated with theory memes and collected records of evidence to initiate the agent based simulation. This provides a first capability to rapidly ingest collected information and preprocess the information within very short time frames.

In addition, this study attempted to demonstrate or document the adequacy of disparate, multi-scale, and dynamic data sources that can validate the causal patterns and the predictive insights extracted from CAS-based models for social or economic systems. The classified data sources concerning those areas suffering the greatest regional instability are essential to the construction of synthetic populations. Unclassified sources are adequate for demonstrating certain principles and methodologies and future technical issues. However, the new simulation systems will create demands for survey and data collection in a fashion similar to the Army’s Human Terrain collection efforts and the Strategic Command’s Human Network Attack capabilities.

In the course of this study, a large number of sources were assembled to acquire information in support of the flexible simulation model. From this study, we have learned that recent government efforts coordinated through the National Geospatial-Intelligence Agency have improved the granularity and coverage of the data required by many social theory simulations. Because motivations and intent of either individuals or small groups are difficult to collect in real-time, databases of populations’ spatial distributions and behaviors require demographic experts to apply data fusion, knowledge discovery, and demographic modeling techniques to build synthetic populations. Fortunately, these techniques exist, and active federal research programs are further developing the techniques. Validation studies on denied areas of the world are continuing to evolve, but a variety of modeling techniques is providing new access to social behaviors in these portions of the world.

ACKNOWLEDGMENT

This research was performed at the Oak Ridge National Laboratory, managed by UT-Battelle, LLC for the U.S. Department of Energy under Contract DE-AC05-00OR22725. The United States Government retains, and the publisher by accepting the article for publication, acknowledges that the United States Government retains, a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes.

The research is part of a broader effort called "Actionable Capabilities for Social and Economic Systems" funded by DARPA [3].

REFERENCES