Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 36 (2014) 198 – 203

Procedia Computer Science

Complex Adaptive Systems, Publication 4
Cihan H. Dagli, Editor in Chief
Conference Organized by Missouri University of Science and Technology
2014-Philadelphia, PA

Modeling Complexity in Multi-modal Adaptive Survey Systems

Fred Highland*

[a] *Lockheed Martin, PO Box 36, New Midway, MD, 21775, USA*

Abstract

Modern survey data collection systems must balance cost and quality while supporting multiple response modes (paper, internet, telephone and personal interview) and addressing unpredictable respondent behavior. The next generation of survey systems utilizes adaptive methods to address these issues, but this may affect system behavior and introduces new issues. The paper discusses the development of a system model to analyze system behavior, determine the level of complexity present, define the conditions under which complex behaviors occur and explore approaches to manage complexity. The model, built in NetLogo, uses an agent representation for control, response mode management and the respondent. It not only represents control logic, particularly survey strategies, but also realistic stochastic respondent demographics and external influences which are independent of the system. The paper frames the problem as a potential complex adaptive system, discusses the approach and modeling of the system and reviews analysis of the model to date and its impacts on system design. Preliminary results indicate that basic system behavior is complicated (according to the Cynefin framework) but external influences can introduce unpredictable behaviors that make the system complex, requiring careful management in order to achieve system objectives.

Peer-review under responsibility of scientific committee of Missouri University of Science and Technology

Keywords: complexity, survey, census, agent based modeling (ABM), NetLogo, external influences

* Corresponding author. Tel.: +1-301-471-4685.
E-mail address: fred.highland@lmco.com

Peer-review under responsibility of scientific committee of Missouri University of Science and Technology
doi:10.1016/j.procs.2014.09.079

1. Introduction

Statistical agencies are under increasing pressure to reduce costs and improve efficiency. Growing costs for survey operations combined with declining budgets and survey response rates dictate the need to find ways to be more efficient and effective, as well as to reduce costs. Part of the solution is to expand data collection methods to include multiple response modes (e.g., paper, internet, telephone and personal interview). This provides new options to respondents, decreasing their response burden and potentially improving response rates as well as reducing costs of collection. Multi-modal data collection has been used successfully in many censuses and surveys. To make further improvements in operational efficiency, Adaptive Survey Design[1,2] has been proposed. Adaptive Survey Design incorporates methods to monitor response data and paradata (data about the collection process) to dynamically adapt collection strategies to optimize quality and/or cost. The combination of multi-modal data collection and adaptive survey design methods dramatically raises the complexity of data capture systems, raising questions about whether such systems will exhibit complex or chaotic behavior, what that behavior is and how it can be controlled.

This work explores survey data collection as a complex adaptive system. In terms of commonly used definitions for complex adaptive systems[3]:

- Uses simple components or agents - While the components may not be simple in their implementation, their behaviors are. Each element has the simple task of requesting and receiving survey responses from respondents. The respondents are also components that react to the requests applying their own behaviors and reactions to external influences.
- Exhibits nonlinear interactions among components - The interactions within the system are ultimately determined by the behaviors of the respondents. Respondent behavior is determined by a probability distribution derived from the interactions of respondent characteristics, response requests and external influences that generally follow time-varying exponentially-distributed patterns. The implementation of other components is generally based on workflow and service-oriented architecture orchestration approaches that involve work queuing which is also inherently exponential in nature.
- No central control - The control element of the system is not strictly in control but only directs the other components to solicit inputs from respondents. Respondent behavior ultimately determines system behavior.
- Emergent behaviors - The objective of this work is to determine if emergent behaviors are exhibited and understand their characteristics. As the system behavior is largely determined by inherently sociocultural respondent actions combined with adaptive methods, it is expected that some self-organization and complicated dynamics will be exhibited. The adaptive survey systems have an inherent adaptive component that represents a form of self-organization and learning. By dynamically adjusting response strategies, they learn to optimize results based on the characteristics and actions of respondents.

The remainder of this paper will discuss the development of a model to understand and predict the complexity aspects of multi-modal adaptive survey data collection systems. It will begin with an explanation of adaptive survey design concepts followed by a description of the design and construction of a model to assess the behavior of the systems. It will conclude with a discussion of results observed to date and plans for future work.

2. Adaptive Survey Design

Survey data collection has traditionally used a static collection design. The process begins by defining the data to be collected and identifying the population to collect the data from. Then an analysis of the characteristics of the population and the available collection modes is performed to determine the optimum plan to collect the data. Optimality may be defined by criteria such as cost, quality, time or other factors. Once the optimum strategy is defined, it is executed to collect the data. If issues arise during the collection process, they are addressed by survey operations in an ad hoc manner. This can lead to cost overruns, delays in the result delivery and even quality issues.

The major deficiency of the traditional survey design approach is the reliance on historical or administrative data as the sole basis for estimates and its inflexibility during execution of the survey. Historical data may not represent the current characteristics of the survey population, resulting in incorrect response expectations. In addition,

unforeseen political, economic and social events may influence respondent behavior during the survey, further adding unpredictability to the process.

Adaptive Survey Design[1,2] has been developed to address the deficiencies of traditional survey design approaches. It begins by defining a set of possible survey strategies (variations on the application of the available collection modes) that can be applied to different subsets of the survey population. It then frames the survey process as a constrained optimization problem:

$$\max_p Q(p) \quad given \quad C(p) \le C_{\max} \tag{1}$$

or

$$\min_p C(p) \quad given \quad Q(p) \ge Q_{\min} \tag{2}$$

Where p represents a matrix of survey allocation probabilities, $Q(p)$ represents the quality resulting from those allocations, $C(p)$ represents the cost of applying those allocations, C_{max} represents the budget for a survey, and Q_{min} the minimum quality requirements. Typically, equation 1 is used for surveys to provide the best quality results from a subset of the total population under limited costs. Equation 2 can be used for censuses where the objective is to survey the entire population to a required level of quality while minimizing the cost.

The optimization process is then applied during survey operation, allowing collection methods to be adapted to current survey characteristics to ensure an optimal result. It also augments the optimization process with data collected during the survey including respondent data and paradata (data about the survey process). This data can be used to correct predictions during collection and adapt to changes in population characteristics that were not available from historical data or are the subject of external influences that were not anticipated.

3. Multi-modal Adaptive Survey System Complexity Model

To understand the complexity characteristics of multi-modal adaptive survey systems, a model was created to simulate their behavior. The model uses agent-based modeling (ABM) approaches with independent agents representing the major elements of the system. Following concepts derived from the Monterey Phoenix work[4,5], the model frames the problem using the major behavioral and environmental components rather than the system design. A graphical representation of the model is shown in Figure 1. The components of the model and their role are:

- **Multi-modal Operations Control System (MOCS)** – provides the overall control for the model and incorporates Adaptive Survey Design decisions. It loads and maintains the survey strategies, initializes the Mode, Respondent and Influence agents and invokes the agents at each time cycle to implement their behavior.
- **Mode Agents** – implement the behaviors of all response modes as directed by Survey Strategies. This includes contacting the respondents and recording their response. A mode agent instance is created for each mode supported (Internet, Paper, Telephone, Field and, in the future, Marketing and Administrative Records (Adm Rec)). Mode Agents implement all forms of response processing behavior for a mode based on survey strategies to contact respondents and record responses.
- **Respondent Agents** – implement the behavior of respondents. There are two behaviors currently implemented: receiving a contact requesting a response by a specified mode and stochastically deciding to respond based respondent mode behavior and influences. Respondents are contacted on each time tick as appropriate until they have responded.
- **Influence Agents** – trigger influence events observed by Respondent Agents to modify response behavior.

Note that in a traditional system design, Respondent and Influence Agents would be external component that may only be represented in data definitions.

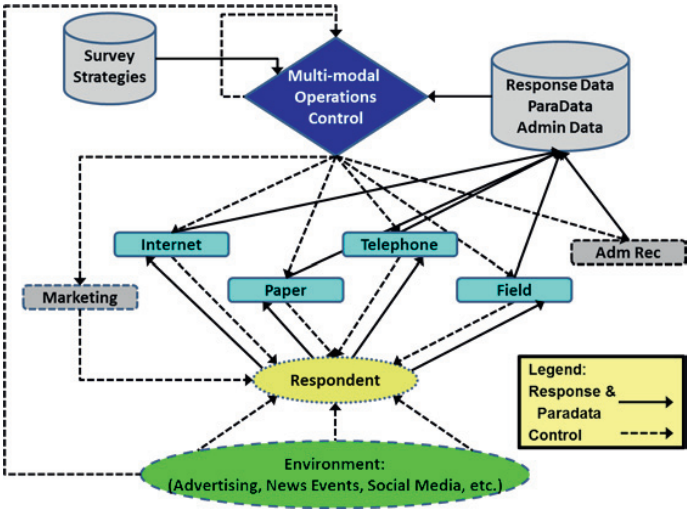

Figure 1 Model for Multi-modal Adaptive Survey Complexity Simulation

The fundamental model used by the system to provide stochastic response behavior is the total response propensity of a respondent. This is defined as:

$$\rho(x,s,t) = \rho_{Mode}\rho_{ModeTime}(t-t_0)\prod_{Influence=1}^{N}\phi_{Influence}\omega_{Influence}(t-t_{0Influence}) \quad (3)$$

Where

- ρ - the response propensity for a respondent with characteristics x applying survey strategy s at time t.
- x - a set of characteristics defined as relations on demographic values assigned to the respondent
- s - a survey strategy
- t - the current time in days since the start of the survey
- ρ_{Mode} - the propensity of a respondent to respond using a mode based on the respondent's demographics. This is typically determined by logistic regression on the historical demographic parameters. It is computed prior to the simulation for efficiency, but it can be recomputed during the simulation if needed or if paradata is used in the computation.
- $\rho_{ModeTime}$ - the probability that a respondent will respond to a mode given the time since the response was requested (t_0). For generality, the simulation defines this as a Gamma distribution on t-t_0 but other methods may be used.
- t_0 - time a contact was made for the mode
- $\phi_{Influence}$ - the effect of an external influence on the response. In order to support positive and negative influences, the range of the influence effect is [0.5,2] where values < 1 are negative influences and values > 1 are positive influences.
- N - the total number of external influences, both positive and negative.
- $\omega_{Influence}$ - the weight of the influence from the Influence Agent. This is a dynamic quantity that may change over time. For generality, the simulation defines this as a Gamma distribution on t-$t_{0Influence}$.
- $t_{0Influence}$ - the time an influence takes effect.

This representation of response propensity was chosen to incorporate the base propensity of the respondent based on demographic data, a time-varying component representing the likelihood of response in the context of other

events in the respondent's life and the time-varying impact of external influences. The response propensity is computed for each non-responding respondent for each time cycle for all modes that have made contact. A random number is then generated and used to determine if a response is to be made.

The simulation is implemented in NetLogo[6]. NetLogo was designed to support agent-based simulations and has a number of features that simplify the task. MOCS is implemented as the NetLogo observer as it is not really an agent. Its task is to initialize the simulation and invoke the other agents for each time cycle. The other agent types are all special instances of "turtles" with instances for each agent type that implement specific behavior. The Mode agents provide functionality to contact respondents and collect information from them. Respondents process response requests and influences and are represented as separate instances per population unit (respondents with like demographics) to allow retention of unique demographics, status and state. To visualize the results, the NetLogo Interface tools were used to specify input datasets and generate graphics to visualize the results. Geographic data was obtained from the US Census Bureau TIGER Products[7] and displayed using the NetLogo GIS extension. An example of the Interface is shown in Figure 2. It features a color-coded map of the US used to indicate response mode by population unit. Graphs of cumulative responses and response rates by day and mode are shown along with the timeframes for strategy applications.

The typical simulation will contain 15,000 respondents (population units). To create realistic test data without getting into issues of Personally Identifiable Information (PII) and violating US Title 13 which protects the confidentiality of collected Census data, a NetLogo synthetic data generator that uses demographic distributions from the US Census Bureau's American FactFinder system[8] was also developed.

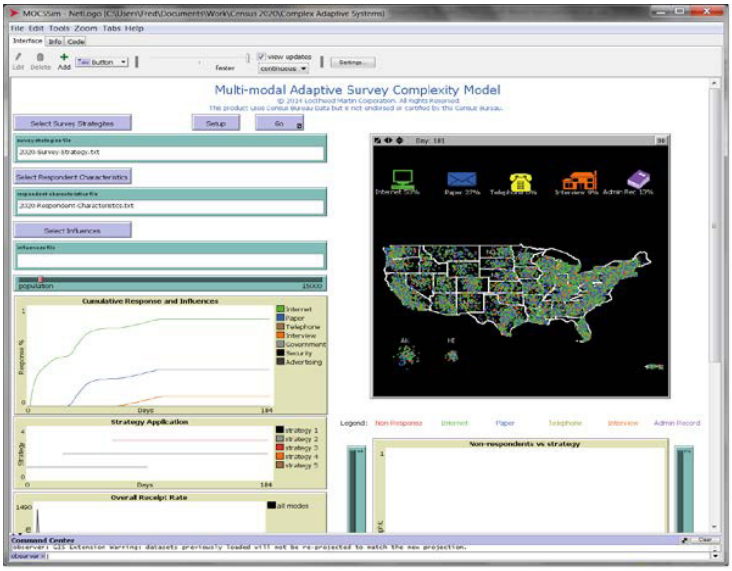

Figure 2 Multi-modal Adaptive Survey Complexity Simulation User Interface

4. Results and Future Direction

Using the Cynefin definition of complex adaptive systems[9] as a measure, it is clear that multi-modal adaptive survey systems are at least complicated. The equations that define the behavior of the different modes and respondents can be accurately defined using modern statistical methods. However, when combined together both time overlaps and behavior interdependencies make the results difficult to predict without extensive simulations. It is also obvious that the system's behavior will enter the realm of complexity when external influences are considered. While the effect of external influences could, in theory, be analyzed and included in the modeling, it is the nature of external influences that their occurrence and effects are not known beforehand. They may therefore create conditions requiring actions of an exploratory nature based on experience, which are characteristic reactions

to complex behavior. The inherently complicated nature of these systems combined with the dynamics introduced by adaptive methods suggests that the system may become chaotic if the adaptive adjustments are too frequent, as has been seen in oscillation effects in control systems[10]. This has not been observed in experiments to date, but as the strategy sets become more complex and volumes of population units get larger, it is a possibility.

The use of NetLogo as a basis for implementation has proven to be an effective approach. The agent-based simulation features, the interactive nature of the language, a wide range of available examples and a simple but effective user interface have greatly simplified development and allowed experimentation. The user interface system is limited and capability and complex representations can be difficult or impossible to create. Also, the list-based nature of NetLogo's data representation, while flexible and efficient, can be difficult to understand, as extracting data as series of list operations can be obscure. The interpretive nature of the language can provide advantages in representing the problem but also limits performance to the order of 100,000 objects (turtles) depending on the nature of the application.

As part of the ongoing work in understanding multi-modal adaptive survey systems, plans include enhancing the models with more complex survey strategies that are being considered for use in future survey operations[11,12,13]. The system could also be enhanced with an adaptive decision component to perform simple optimization of strategies. This is a potentially complex area requiring compute-intensive algorithms, and it is not the intent to develop a full-scale optimization capability. Simplistic methods and predefined optimization scenarios are being considered, as they are more practical for the need.

The simulation models developed here will be used for research in multi-modal adaptive survey system characteristics so that future survey systems can address complex environments involving millions of respondents and unknown influences and provide a basis for specifying the design and interfaces for those future systems.

References

1. Wagner, J. Adaptive Survey Design to Reduce Nonresponse Bias. Dissertation. University of Michigan. http://deepblue.lib.umich.edu/bitstream/handle/2027.42/60831/jameswag_1.pdf?sequence=1 . 2008.
2. Schouten. B, Calinescu, M, and Luiten, A. (2011). "Optimizing quality of response through adaptive survey designs", Statistics Netherlands, Henri Faasdreef 312 2492 JP The Hague. http://www.cbs.nl/NR/rdonlyres/2D62BF4A-6783-4AC4-8E45-12EF20C6675C/0/2011x1018.pdf
3. Mitchell, M. Introduction to Complexity. Santa Fe Institute. http://s3.amazonaws.com/complexityexplorer/IntroToComplexity/Unit1Slides.pdf. 2013.
4. Auguston, M. Software Architecture Built from Behavior Models. ACM SIGSOFT Software Engineering Notes, 34:5. 2009.
5. Giammarco, K., Auguston, M. Well, you didn't say *not* to! A Formal Systems Engineering Approach to Teaching an Unruly Architecture Good Behavior. Complex Adaptive Systems 2013. Elsevier B.V. 2013.
6. Wilensky, U. NetLogo. http://ccl.northwestern.edu/netlogo/. Center for Connected Learning and Computer-Based Modeling, Northwestern University, Evanston, IL. 1999.
7. US Census Bureau. TIGER Products TIGER/Line Shape Files 2013. http://www.census.gov/geo/maps-data/data/tiger.html . 2014.
8. US Census Bureau. American FactFinder. http://factfinder2.census.gov/faces/nav/jsf/pages/download_center.xhtml. 2014.
9. Snowden, David. "Multi-ontology sense making – a new simplicity in decision making". *Management Today Yearbook*, 2005. http://cognitive-edge.com/uploads/articles/40_Multi-ontology_sense_makingv2_May05.pdf . 2005.
10. Andrievskii, B., Fradkov, A. Control of Chaos: Methods and Applications. II. Applications. AUTOMATION AND REMOTE CONTROL Vol. 65 No. 4 2004. ftp://ftp.elet.polimi.it/users/Carlo.Piccardi/VarieCda/Articoli/CdA-Art-ControlloSincro-6.pdf . 2004
11. Matthews, B., Davis, M.C., Tancreto, J.G, Zelenak, M.F., Ruiter, M. 2011 American Community Survey Internet Tests: Results from the Second Test in November 2011. 2012 American Community Survey Research and Evaluation Report Memorandum Series #ACS12-RER-21-R1. http://www.census.gov/acs/www/Downloads/library/2012/2012_Matthews_01.pdf . 2012.
12. Tancreto, J. G., Zelenak, M. F., Davis, M. C., Ruiter, M., and Matthews, B. L. "2011 American Community Survey Internet Tests: Results from First Test in April 2011," 2012 American Community Survey Research and Evaluation Report Memorandum Series #ACS12-RER-13-R2, June 25, 2012 2012.http://www.census.gov/acs/www/Downloads/library/2012/2012_Tancreto_01.pdf .2012.
13. US Census Bureau. Census Barriers, Attitudes, and Motivators Survey II Final Report. 2010 Census Integrated Communications Research Memoranda Series. http://www.census.gov/2010census/partners/pdf/CBAMS_II_Final_Report.pdf . 2011.