Comput Math Organ Theory
DOI 10.1007/s10588-012-9114-1

Case-based modeling and the SACS Toolkit: a mathematical outline

Brian Castellani · Rajeev Rajaram

Abstract Researchers in the social sciences currently employ a variety of mathematical/computational models for studying complex systems. Despite the diversity of these models, the majority can be grouped into one of three types: agent (rule-based) modeling, dynamical (equation-based) modeling and statistical (aggregate-based) modeling. The purpose of the current paper is to offer a fourth type: case-based modeling. To do so, we review the SACS Toolkit: a new method for quantitatively modeling complex social systems, based on a case-based, computational approach to data analysis. The SACS Toolkit is comprised of three main components: a theoretical blueprint of the major components of a complex system (*social complexity theory*); a set of case-based instructions for modeling complex systems from the ground up (*assemblage*); and a recommended list of case-friendly computational modeling techniques (*case-based toolset*). Developed as a variation on Byrne (in Sage Handbook of Case-Based Methods, pp. 260–268, 2009), the SACS Toolkit models a complex system as a set of k-dimensional vectors (cases), which it compares and contrasts, and then condenses and clusters to create a low-dimensional model (map) of a complex system's structure and dynamics over time/space. The assembled nature of the SACS Toolkit is its primary strength. While grounded in a defined mathematical framework, the SACS Toolkit is methodologically open-ended and therefore adaptable and amenable, allowing researchers to employ and bring together a wide variety of modeling techniques. Researchers can even develop and modify the SACS Toolkit for their own purposes. The other strength of the SACS Toolkit, which makes it a very effective technique for modeling large databases, is its ability to compress data matrices while preserving the most important aspects of a complex system's structure and

B. Castellani (✉)
Dept. of Sociology, Kent State University, Ashtabula, OH 44004, USA
e-mail: bcastel3@kent.edu

R. Rajaram
Dept. of Mathematical Sciences, Kent State University, Ashtabula, OH 44004, USA

Published online: 20 March 2012

dynamics across time/space. To date, while the SACS Toolkit has been used to study several topics, a mathematical outline of its case-based approach to quantitative analysis (along with a case study) has yet to be written–hence the purpose of the current paper.

Keywords Complex social systems · Case-based method · Mathematical modeling · Computational modeling · Sociological method

1 Introduction

Researchers in the social sciences currently employ a variety of mathematical/computational models for studying complex systems. Despite the diversity of these models, the majority can be grouped into one of three types: agent (rule-based) modeling, dynamical (equation-based) modeling and statistical (aggregate-based) modeling. The purpose of the current paper is to offer a fourth type: case-based modeling (Castellani and Hafferty 2009; Castellani et al. 2012). Case-based modeling (although traditionally not computational) is an established technique in the social sciences, used for conducting in-depth, idiographic, comparative analyses of cases and their variable-based configurations (Rihoux and Ragin 2009). Given its orientation, case-based modeling is regularly used as a viable alternative to variable-based statistics (Ragin 2008). Only recently, however (as in the last few years) have researchers begun to explore its potential for modeling complex systems, particularly with large databases and as a computational technique (Byrne 2009; Castellani et al. 2012; Janasik et al. 2008; Uprichard 2009). The premise for this new line of inquiry, expressed by Byrne (2009), is that cases are the methodological equivalent of complex systems.

One such example of a case-based model designed for studying complex systems, which we review here, is the SACS Toolkit (Castellani and Hafferty 2009). The SACS Toolkit is a case-based, mixed-method, system-clustering, data-compressing, theoretically-driven toolkit for modeling complex social systems. It is comprised of three main components: a theoretical blueprint for studying complex systems (*social complexity theory*); a set of case-based instructions for modeling complex systems from the ground up (*assemblage*); and a recommend list of case-friendly modeling techniques (*case-based toolset*). The SACS Toolkit is a variation on Byrne's (2009) general premise regarding the link between cases and complex systems. For the SACS Toolkit, case-based modeling is the study of a complex system as a set of n-dimensional vectors (cases), which researchers compare and contrast, and then condense and cluster to create a low-dimensional model (map) of a complex system's structure and dynamics over time/space. Because the SACS Toolkit is, in part, a data-compression technique that preserves the most important aspects of a complex system's structure and dynamics over time, it works very well with databases comprised of a large number of complex, multi-dimensional, multi-level (and ultimately, longitudinal) variables. It is important to note, however, before proceeding, that the act of compression is different from reduction or simplification. Compression maintains complexity, creating low-dimensional maps that can be "dimensionally inflated"

as needed; reduction or simplification, in contrast, is a nomothetic technique, seeking the simplest explanation possible.

The SACS Toolkit is also versatile and consolidating. The strength, utility, and flexibility of the SACS Toolkit comes from the manner in which it is, mathematically speaking, put together. The SACS Toolkit emerges out of the assemblage of a set of existing theoretical, mathematical and methodological techniques and fields of inquiry. The "assembled" quality of the SACS Toolkit, however, is its strength. While it is grounded in a highly organized and well defined mathematical framework, with key theoretical concepts and their relations, it is simultaneously open-ended and therefore adaptable and amenable, allowing researchers to integrate into it many of their own computational, mathematical and statistical methods. Researchers can even develop and modify the SACS Toolkit for their own purposes.

To date, a basic (non-mathematical) introduction to the SACS Toolkit has been written (Castellani and Hafferty 2009); and the SACS Toolkit has been (or is currently being) used to study several topics, including medical professionalism (Castellani and Hafferty 2009, 2010), medical education, allostatic load, MRSA (Methicillin-resistant Staphylococcus aureus) transmission, school systems, and community health (Castellani et al. 2012). However, a mathematical outline of its case-based approach to quantitative analysis (along with a case study) has yet to be written—hence the purpose of the current paper, which is organized as follows.

For this paper, we outline, point by point/concept by concept, the SACS Toolkit's basic mathematical framework. At each point along the way we do two things: (a) define the concept and (b) illustrate its purpose by applying it to our case study. As a final note, we have kept the mathematics and the case study in this paper simple to facilitate the widest possible understanding, with the idea that future papers (by ourselves and other researchers) will provide detailed, case-driven reviews of the more technical aspects of the SACS Toolkit, such as how it models complex systems across time/space.

The SACS Toolkit: A Mathematical Outline

2 Cases c_i as k-dimensional vectors

We begin our review of the SACS Toolkit with five opening points:

(1) For the SACS Toolkit, case-based modeling is the study of a complex system S as a set of cases c_i such that:

$$S = \{c_i : c_i \text{ is a case relevant to the system under study}\}. \quad (1)$$

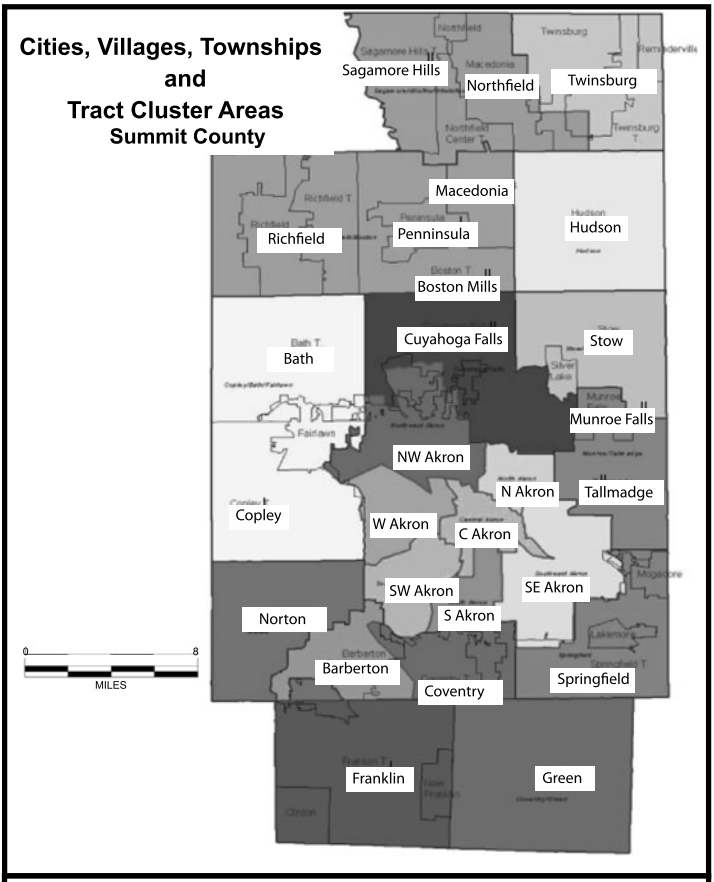
(2) At minimum, S is comprised of one case c_i.
(3) While there is no predefined limit to the maximum number of cases that can be included in the study of a complex system, practically speaking the upper limit will be bounded, based on the particular set of cases identified for study—which is always an empirical issue.
(4) We denote the number of cases being studied by n.

(5) Each case c_i in S is a k dimensional row vector $c_i = [x_{i1}, \ldots, x_{ik}]$, where each x_{ij} represents a measurement on one of the variables being used to model a complex system.

To illustrate these five points empirically, we turn to our case study.

2.1 Case study: community health

The case study for our paper is a simplified and abbreviated version of a community health science study we recently completed (Castellani et al. 2012), examining the health and wellbeing of 20 communities in Summit County, Ohio, USA. As shown in Fig. 1, for our case study, S is defined as the set of 20 communities in Summit



Source: This map was retrieved from http://healthysummit.org/QOL/QOL_Maps.cfm on the 29th of November, 2011. It is a public document provided by the Health Summit 2010 website. SOURCE: Federation for Community Planningvand Northern Ohio Delta & Information Services (NODIS) Maxine Goodman Levin College of Urban Affairs, Cleveland State University, USA, April 2003.

Fig. 1 Map of 20 communities in *Summit*

County, USA. For practical purposes, we call this complex system of communities *Summit*.

In terms of getting to know this case study, as shown in Fig. 1, *Summit* has a major urban center, the City of Akron, which is comprised of a set of urban communities, including Barberton, which is a smaller city to Akron's south. Moving outward from Akron there is a set of first-tier suburban communities, mostly working and middle class. There are also communities on the edges of Summit County that are upper-class and affluent, such as Hudson, Peninsula and Bath. And, finally, at the bottom of *Summit*, there are a few semi-rural communities, such as Green. The goal of our case study is to model these communities as a complex system to understand better the underlying structure and dynamics of their respective health and wellbeing.

Following our first five points, each of the $(n = 20)$ communities c_i in *Summit* is a k dimensional row vector $c_i = [x_{i1}, \ldots, x_{ik}]$, where each x_{ij} represents a measurement on one of the variables we will use to model *Summit*, which we will define next.[1]

3 Cases as social practices P_i and environmental forces E_i

The SACS Toolkit's definition of cases as complex systems is not only grounded in case-based modeling. It is also grounded in the theoretical tradition known as *practice theory* (Ahearn 2001; Dreyfus and Rabinow 1983; King 2004)—which comes from the work of Michel Foucault (Foucault 1980; Giddens 1980; Bourdieu 1990). As a type of practice theory, the SACS Toolkit has a distinct view about the types of factors (variables, etc.) that make up the k dimensional row vectors for the c_i comprising S. This distinct view is outlined in the following seven points:

(1) The most fundamental unit of sociological analysis is social practice.
(2) A complex system is comprised of a set of social practices W_s, which consists of social practices denoted by P_i.
(3) As a set of social practices W_s, a complex system S is externally impacted by a set of environmental forces E_s, which consists of environmental forces denoted by E_i.
(4) A complex system S self-organizes around and emerges out of the coupling of social practices P_i and environmental forces E_i.
(5) The k-dimensional vector for each case c_i in S is comprised of a set of measurements on key social practices P_i and environmental forces E_i.
(6) As a k-dimensional row vector $c_i = [x_{i1}, \ldots, x_{ik}]$ (comprised of a set of measurements on W_s and E_s), each case c_i constitutes one of the possible ways a complex system S can be practiced.
(7) A complex system S is therefore a set of cases $\{c_i\}$, with each case c_i constituting one of its possible ways of practice, based on the coupling of W_s and E_s.

Let us explore these points in more detail.

[1] For those interested in an detailed explanation of the SACS Toolkit in application to our case study, see Castellani et al. (2012).

4 Social practices P_i and environmental forces E_i

The variables x_{ij} used to construct the dimensions of the vectors in S are one of two types: social practices P_i or environmental forces E_i. In Castellani and Hafferty (2009) the concept of social practice is addressed in detail. Here, a brief explanation is provided. As a type of practice theory, the SACS Toolkit uses the concept of social practice to move past several key dualisms in social scientific inquiry: structure/agency, objective/subjective, quantitative/qualitative, micro/macro, and bottom-up/top-down. Social practice is defined as any pattern of social organization that emerges out of, and allows for, the intersection of symbolic interaction and social agency (Castellani and Hafferty 2009, p. 38). Examples range from the social practices people employ to care for themselves, such as health behaviors, to those used in caring for others, such as health care or public health.

Regardless of its level of complexity, a social practice is comprised of five key components: interaction, social agents, communication, coupling and social knowing. Formally, we define a social practice as a five-tuple, noting it as follows:

$$P = (A, I, L, C, K). \tag{2}$$

In this notation, $A = \textit{social agent}$ is defined as the means or instrument by which a social practice achieves its results; social agents are those things that produce or are capable of producing an effect or outcome. A social agent can be a person, computer, organization, society, etc. $I = \textit{Interactions}$ is defined as the key relationships within a social practice. It is a kind of exchange that occurs as two or more agents have an effect upon one another; a mutual or reciprocal action. $L = \textit{Communication}$ is defined as the act of connecting and transferring information and knowledge within or amongst the agents in a social practice. $C = \textit{Coupling}$ is defined as the agent-based linking of two or more social practices through a common intermediary (typically something that is considered part of the social practices) to form the complex system of which they are a part. Coupling also links social practices to biological and physical practices/systems. Finally, $K = \textit{Social Knowing}$ is defined as the qualitative awareness, intelligence and understanding of a social practice through human agency. Social practices are not just a set of agent-based rules.

For practical purposes, the empirical measurement of P_i can vary greatly in granularity. In some instances, researchers may find it necessary to provide great empirical detail about each P_i; in other instances, each P_i may be a single measure. It all depends upon the level of complexity required for a particular study of some complex system S.

The set of social practices out of which S emerges is referred to as the *web of social practices* W_s. We write this as follows:

$$W_s = \{P_i : P_i \text{ is a social practice relevant to the system of study}\}. \tag{3}$$

Environmental forces are those external forces E_i (be they whatever, an external system, event, etc.) that influence a complex system's web of social practices W_s. The environmental forces reside outside W_s and are independent of the social practices themselves. We write the set of all environmental forces as follows:

$$E_s = \{E_i : E_i \text{ is an environmental force relevant to the system of study}\}. \tag{4}$$

Practically speaking, the vectors c_i in S are multi-dimensional, often comprised of measurements on a large number of W_s and E_s, and usually across multiple levels of scale. At minimum, however, the vectors c_i in S must contain a measurement on at least one P_i.

Motivated by the possibility that two or more social practices (along with one or more environmental forces) will become the actual variables used to model a complex system, we define a finite partition of W_s or E_s as follows. A finite partition of W_s is a collection of non-empty subsets $\{O_i\}_{i=1}^{p}$ of W_s satisfying the following properties:

$$\begin{aligned} &\text{(a) } O_i \cap O_j = \emptyset \quad \forall i \neq j. \\ &\text{(b) } \bigcup_{i=1}^{p} O_i = W_s. \end{aligned} \tag{5}$$

The same definition of the partition of W_s applies to E_s to describe a single environmental force or a collection of forces.

Now that we have a basic understanding of the variables x_{ij} (social practices P_i and environmental forces E_i) used to construct the dimensions of the vectors in S, we turn to our case study to empirically ground this basic understanding.

4.1 Case study

In terms of our case study, we used five variables to construct the vectors for our cases (communities): (1) urban sprawl, (2) household income, (3) education level, (4) border sharing and (5) years of life lost.

As the P_i in W_s, household income, educational level, years of life lost and border sharing are each a five-tuple; and, depending upon the level of complexity sought, can be measured as such. Let us consider, for example, border sharing. *Border sharing* tells us which communities share a border with one another. This measure is important for exploring how communities, based on their geo-socio-political position, influence the vector configuration of other communities, specifically their health outcomes. In border sharing, the A could be all 20 communities in *Summit*. I could be any or all key interactions amongst communities that share a border with one another. L could be all internal or external acts of communication regarding the political, economic, cultural and geographical links amongst communities. These communications could be conducted by a host of agents: citizens, political leaders, newspapers, businesses, etc. C, amongst other things, could be how the various border communities couple (based on similarities in their respective vector configurations) to influence the health and wellbeing of *Summit*. And, finally, K could be the self-awareness of a community about what it needs to do to improve its resident's health.

However, while the complexity of each P_i in W_s and our E_s can be fully measured, to keep our case study simple, each P_i in *Summit* was defined as a single measure. *Household income* was defined as the median household income for each of the 20 communities. *Educational level* was defined as the number of persons 25+ Years in each community with a high school diploma. *Years of Life Lost* was defined as the sum of the differences between the age at death and the life expectancy at age of death

for each death occurring between 1990–1998 in each community due to all causes, divided by the number of deaths due to all causes within each community. Finally, while *Years of Life Lost* is one of the P_i in W_s, it also functions as the outcome variable for our study: a way of comparing the communities in our study (think case-based modeling) based on a key health issue.

Urban sprawl represents the E_s in our model, and is defined as the unplanned out-migration (flow) of relatively low-density development into the suburban and rural tiers surrounding an urban area in a particular region. We measured urban sprawl using an agent-based model, which we discuss later.

Now that we have a basic understanding of the W_s and E_s in our case study, the next step is to understand how they go together to form our communities (cases), which takes us to the issue of *coupling*.

5 Cases as the coupling of W_s and E_s

While the k dimensions of each vector $c_i \in S$ are discrete measurements on W_s and E_s, the SACS Toolkit does not consider these k dimensions devoid of relationship. Remember, social practices are agent-based forms of social organization. As such, the SACS Toolkit assumes that, through agency, the k dimensions of c_i are interconnected and relational. The formal term for these interconnected, agent-based relations (as explained above in the definition of social practice) is *coupling*. In fact, it is out of the agent-based coupling of W_s and E_s that the different vectors $c_i \in S$ emerge. In other words, each vector $c_i \in S$ constitutes one of the possible ways that a complex system (as a set of W_s and E_s) can be practiced/expressed. In turn, it is out of the case-based coupling of W_s and E_s that the complex causal structure and dynamics of S emerge.

In case-based modeling terms, the particular way that W_s and E_s couple to form the k dimensions of each c_i in S is called that case's *vector configuration*. To understand these vector configurations, we begin with a review of the database they form.

5.1 Cases as databases

Because S consists of n cases $\{c_i\}_{i=1}^{n}$, and each case c_i has a vector configuration of k dimensions, it is natural to represent S, at least initially and at its most basic, in the form of a data matrix D as follows:

$$D = \begin{bmatrix} c_1 \\ \vdots \\ c_n \end{bmatrix} = \begin{bmatrix} x_{11} & \dots & x_{1k} \\ \vdots & \ddots & \vdots \\ x_{n1} & \dots & x_{nk} \end{bmatrix}. \tag{6}$$

In the notation above, the n rows in D represent the set of cases $\{c_i\}$ in S, and the k columns represent the measurements on some finite partition $\bigcup_{i=1}^{p} O_i$ of W_s and E_s as defined in Eq. (5) that couple to form the vector configuration for each c_i. The complexity of the system is exemplified by the fact that any element of the partition can range from the finest (where all elements are singleton sets) to the coarsest

(e.g. where the partition has a single element, namely, a set consisting of all social practices, P_i). However, we envision a typical partition to be somewhere in between the two extremes described above i.e., a partition that has some elements O_i that are possibly singleton sets and some that are not. In other words, although the mathematical possibilities are many, care must be taken to select meaningful elements in the partition.

Putting the above together, we can summarize D as shown in (6), where each $c_i \in D$ is a case-based k dimensional row vector $c_i = [x_{i1}, \ldots, x_{ik}]$ comprised of a set of measurements on W_s and E_s; and, the k-dimensions of each c_i in S can be thought of as the vertices of a graph, with vertices being non-empty elements of any finite partition of W_s or E_s. As we will see in the upcoming sections, it is the goal of the SACS toolkit to concretely establish relationships between the elements O_i of the partition. A relationship between the partitions O_i and O_j can be signified mathematically as a two-tuple (O_i, O_j) and can also be visualized as an edge between the vertices corresponding to O_i and O_j. To demonstrate the empirical utility of these points, we turn to our case study.

5.2 Case study

In terms of our case study of *Summit*, each community ($n = 20$) in D is comprised of a set of measurements on our five key dimensions: urban sprawl, household income, educational level, border sharing and years of life lost. Together, these measurements couple to form the vector configuration for each case (community) in *Summit*. With this database for *Summit* constructed, the next step is to map out its field of relations, which we turn to next.

6 Field of relations R_s

To recall, the purpose of the SACS Toolkit is to model a complex system S as a set of k dimensional vectors (cases c_i), which researchers compare and contrast, and then condense and cluster to create a low-dimensional model (map) of a complex system's structure and dynamics over time/space. The first step of comparing and contrasting the vector configurations in S requires D to be converted into some type of case-based relational matrix. The logic for this conversion is as follows. If each $c_i \in D$ is a case-based k dimensional row vector $c_i = [x_{i1}, \ldots, x_{ik}]$, comprised of a set of measurements on W_s and E_s, then D can be transformed into a set of relational matrices indicating the various pairwise relationships amongst the cases (vector configurations) in D for the purposes of comparing and contrasting cases as different expressions of S. This set of relational sub-matrices is called the *field of relations* R_s.

The field of relations R_s is the heart of the SACS Toolkit, as it provides all the information necessary for assembling a case-based model of S. Specifically, R_s provides a fundamental set of maps for the three main types of relationships (which are, themselves, interdependent and interconnected) that exist amongst the cases $\{c_i\}$ in S. Mathematically, these relations emerge by transforming R_s into three types of relational matrices: (1) the proximity (dissimilarity) matrix $\mathcal{P}$, (2) the adjacency (network) matrix $\mathcal{A}$ and (3) the correlation matrix $\mathcal{L}$. Each of these matrices provides

a particular type of information about the relationships among cases, which can be used to compare and contrast, and then condense and cluster them. Hence we have the following representation of the field of relations R_s as a three-tuple:

$$R_s = (\mathcal{P}, \mathcal{A}, \mathcal{L}). \tag{7}$$

6.1 R_s as proximity (dissimilarity) matrix $\mathcal{P}$

The first relational matrix for R_s is the n by n proximity (or dissimilarity) matrix $\mathcal{P}$ (see (8)). In this case, R_s is represented by an n by n symmetric proximity matrix $\mathcal{P}$ —which can be thought of as a similarity matrix or dissimilarity matrix, depending upon the technique used to partition the data matrix D. As a proximity matrix, the entries on $\mathcal{P}$ indicate the pairwise distances d_{ij} (usually Euclidean) between the cases c_i and c_j from D, based on differences in their respective vector configurations.

$$\mathcal{P} = \begin{bmatrix} d_{11} & \dots & d_{1n} \\ \vdots & \ddots & \vdots \\ d_{n1} & \dots & d_{nn} \end{bmatrix}. \tag{8}$$

The above proximity matrix can be visualized as the edges of a graph $\mathcal{P}_g$, with vertices being the cases $\{c_i\}$, and the edge lengths being the relative distance between them. Such a graph $\mathcal{P}_g$ provides a complete map of the pairwise similarities and dissimilarities amongst cases, which can be used to compare and contrast them.

6.1.1 R_s as adjacency (network) matrix $\mathcal{A}$

The second relational matrix for R_s is the n by n adjacency (network, $\mathcal{A}$) matrix. The $\mathcal{A}$ matrix is typical of the matrices used in the new science of networks (Newman 2010). In this case, R_s is partitioned into all possible choices of two-tuples of cases (c_i, c_j), which can be graphed as a network $\mathcal{A}_g$, where each node constitutes a case, and where each edge constitutes some type of tie (link, connection, social relationship, interaction) amongst these cases. The edges for the nodes in $\mathcal{A}_g$ are defined according to some finite partition of W_s or E_s in D.

$\mathcal{A}$ provides important additional information for R_s. Whereas $\mathcal{P}_g$ provides a complete map of the distances amongst cases, based on differences in their respective vector configurations, $\mathcal{A}_g$ provides a complete map of qualitative ties amongst cases, based on their respective vector configurations. Given that the ties amongst the cases $\{c_i\}$ in S can be measured on a variety of k dimensions (e.g., friendship, kinship, group membership, geographical location, types of interaction, etc.), it is typical for R_s to be partitioned numerous times to create multiple maps $\mathcal{A}_g$.

Also, for the purposes of clarity, it is important to emphasize that, given some finite partition of W_s or E_s in D, the ties in $\mathcal{A}_g$ are not just relationships, but also interactions. As interactions, these ties represent a discrete moment t_i in the dynamic evolution of S, constituting one measurement of the exchange amongst cases; that is, the effect they have on one another.

6.2 R_s as correlation matrix

The third relational matrix for R_s is the k by k correlation matrix $\mathcal{L}$. In this case, R_s is partitioned into all possible choice of pairwise, zero-order correlations between each of the k dimensions in D. In this matrix $\mathcal{L}$, which is symmetric and square, the diagonal elements are equal to 1. The utility of $\mathcal{L}$ is that it provides a complete map of the strength and direction of relationships shared amongst all the k dimensions comprising the vector configurations for the cases $\{c_i\}$ in S. Like $\mathcal{A}$, given that the relationships amongst the variables comprising $\mathcal{L}$ can be studied from a variety of perspectives and therefore partitioned in a variety of ways, it is typical for R_s to be partitioned numerous times to create multiple maps $\mathcal{L}_g$.

Now that we have a basic understanding of the *field of relations* R_s and its three sub-matrices, we turn to our case study to empirically ground these ideas.

6.3 Case study

The field of relations for our case study is comprised of three maps ($\mathcal{P}$, $\mathcal{A}$ and $\mathcal{L}$) which provide us key information about the relationships amongst the communities $\{c_i\}$ in *Summit*.

First, to generate $\mathcal{P}$ we used the conventional statistical software package *SPSS*. Specifically, we used the classification command *hierarchical cluster analysis*, choosing the proximity matrix option with Euclidean distance as our measure. Figure 2 is a snapshot of the proximity matrix for a few of the ($n = 20$) communities in *GPA*. The Euclidean distances in Fig. 2 provide a rough approximation of how conceptually far away one community is from the other, based on differences in their respective vector configurations. In this case, to keep our example simple, border sharing and urban sprawl were not part of the vector configuration analyzed. In Fig. 2, the larger the Euclidean number, the greater the conceptual distance. For example, the vector configuration of Twinsburg is very similar to Northwest Akron, but very different from Hudson.

Next, we generated $\mathcal{A}$. As one example of the relationships amongst the communities that one might study, we focused on border sharing.

Our interest in border sharing follows a trend in the current community health science literature, which argues that communities are best understood "as nodes in

	Euclidean Distance				
Case	1:Sagamore/ Macedonia/N	2:Twinsburg	3:Richfield/ Boston/Pen	4:Hudson	5:Copley/ Bath/Fairlawn
1:Sagamore/Macedonia/N	.000	7534.000	51.057	26004.004	7065.004
2:Twinsburg	7534.000	.000	7483.000	33538.003	14599.002
3:Richfield/Boston/Pen	51.057	7483.000	.000	26055.003	7116.002
4:Hudson	26004.004	33538.003	26055.003	.000	18939.002
5:Copley/Bath/Fairlawn	7065.004	14599.002	7116.002	18939.002	.000
6:Cuyahoga Falls	11212.000	3678.001	11161.000	37216.002	18277.001
7:Stow/Silver Lake	1794.005	5740.001	1743.001	27798.002	8859.001
8:Northwest Akron	5169.004	2365.005	5118.002	31173.001	12234.000
9:Munroe Falls/Tallmad	3441.001	4093.002	3390.000	29445.003	10506.001
10:North Akron	20020.003	12486.007	19969.004	46024.007	27085.006

Fig. 2 Proximity matrix for GPA

		1990 household income	% no highschool 25yrs+ in 1990	Years lost per death 1998
1990 household income	Pearson Correlation	1	-.908**	-.794**
	Sig. (2-tailed)		.000	.000
	N	20	20	20
% no highschool 25yrs+ in 1990	Pearson Correlation	-.908**	1	.752**
	Sig. (2-tailed)	.000		.000
	N	20	20	20
Years lost per death 1998	Pearson Correlation	-.794**	.752**	1
	Sig. (2-tailed)	.000	.000	
	N	20	20	20

**. Correlation is significant at the 0.01 level (2-tailed).

Fig. 3 Correlation matrix for GPA

networks" rather than "discrete and autonomous bounded spatial units" (Cummins et al. 2007, p. 1827). As Cummins et al. state in Cummins et al. (2007): "Studies often ignore issues of spatial autocorrelation and assume that conditions in each locality operate on population health independently of conditions in other areas" (p. 1832). To generate our matrix, we defined $\mathcal{A}$ as follows: $\mathcal{A}$ is a matrix with elements $\mathcal{A}_{ij}$ such that

$$A = \begin{cases} 1 & \text{if communities share a border;} \\ 0 & \text{if otherwise.} \end{cases}$$

The resulting matrix, which was created with the network software package *Pajek*, was non-directed, symmetrical and square, with zeros running along the diagonal.

Finally, we generated $\mathcal{L}$. For $\mathcal{L}$ there were a variety of relationships that could be explored. For example, one of the most commonly explored set of relationships in the community health science literature is the link between household income, educational level and health outcomes, such as years of life lost. As shown in Fig. 3, which we created in SPSS, all three variables, as common sense would suggest, are highly correlated.

With our three matrices constructed, the next step was to condense them to form the *Network of Attracting Clusters*, which we turn to next.

7 Network of attracting clusters $\mathcal{N}$

While the field of relations R_s provides a set of fundamental maps of the various relations between the cases $\{c_i\}$ in S, such a map is rarely useful, particularly when it comes to databases D comprised of a large number of cases $\{c_i\}$ or k dimensions. Needed, instead, are simplified, low-dimensional maps that compress the complexity of R_s, while preserving the most important aspects of its structure and dynamics across time/space. The creation of such a set of simplified maps is the purpose of the *network of attracting clusters* $\mathcal{N}$.

$\mathcal{N}$ is assembled by compressing the three matrices of R_s to create three types of maps: cluster maps, network maps and k dimensional maps. Once assembled, these

three maps, and the network of attracting clusters $\mathcal{N}$ they form—note the words 'network' and 'clusters' in this term—constitute the SACS Toolkit's formal model of a complex system S. Before proceeding to a review of these maps, however, three caveats are necessary:

7.1 A caveat on the mathematics for $\mathcal{N}$

The uniqueness and utility of $\mathcal{N}$ is not found in its mathematical underpinnings; nor are the mathematical ideas upon which it is based new (e.g., clustering, exploring social networks, examining vector configurations, compressing correlation matrices). In fact, the case-based toolset of techniques used to assemble $\mathcal{N}$ is grounded in a well-established and rather extensive literature. For example, as will be explained below, the SACS Toolkit typically compresses the proximity matrix in R_s using two popular algorithms: k-means and the Kohonen self-organizing map (SOM). In turn, the adjacency matrix is compressed using a variety of well-established network measures: centrality, eigenvector centrality, hubs, authorities, cliques, etc. And, the correlation matrix is compressed using such techniques as regression, principal components analysis, agent-based modeling, variable-based modeling, etc. The uniqueness and utility of $\mathcal{N}$ comes, instead, from its integration of these three types of maps, and its consolidation of the techniques needed to data-mine them. Such a 'data integrated' and 'technique consolidating' approach to modeling (at least to our knowledge) does not exist. $\mathcal{N}$ therefore synthesizes the various areas of inquiry in complexity science into a single, unified model.

7.1.1 A caveat on model building

Having made the above caveat, we nonetheless recognize that, while the assemblage of $\mathcal{N}$ involves the creation of all three sets of maps, it is not always what researchers may need. For example, a study may only seek to model $\mathcal{N}$ as a set of cluster maps. Once again, one of the main utilities of the SACS Toolkit is its flexibility.

7.1.2 A caveat on our case study

Lastly, the challenge in this section of the paper is that here, more than anywhere, explaining things by example is highly useful because, in terms of the SACS Toolkit, the construction of $\mathcal{N}$ is the center of the model building process. At the same time, however, given the'assembled' nature of the SACS Toolkit, the danger of driving explanation by example is that it clouds the full flexibility of this technique. Having made these two points, we remind readers that, to facilitate the widest reading, we have kept our case study in this section extremely basic, sufficient only to quickly demonstrate one possible way of using the SACS Toolkit. And, we remind readers that a detailed overview of how we used the SACS Toolkit to model *Summit*, including the creation of the maps in $\mathcal{N}$ can be found in Castellani et al. (2012).

7.2 $\mathcal{N}$ as cluster map

The clustering of $\mathcal{P}$ is based on the methodological argument of Byrne (2009) and like-minded, computationally grounded, case-based researchers (Castellani et al. 2012; Janasik et al. 2008; Rihoux and Ragin 2009; Uprichard 2009). It goes as follows: if the vector configurations for each case c_i in $\mathcal{P}$ constitute a map of the different ways a complex system S can be practiced, then an effective method of compressing $\mathcal{P}$ is to cluster it. Clustering $\mathcal{P}$ is useful because it allows for the identification, mapping and analysis of the most common vector configurations in S across time/space.

At present, there are numerous methods (thousands even) for clustering proximity matrices (see Jain 2009 for a detailed history of cluster analysis). The SACS Toolkit finds two methods particularly useful: the k-means algorithm and the SOM. Following (Sperandio and Coelho 2004), the SACS Toolkit uses these two techniques in combination because their similarities and differences compliment each other well. (Others, however, may find different clustering methods equally useful.)

In terms of similarities, these two techniques share a common approach to data compression and clustering. Both k-means and the SOM can be viewed as vector quantization techniques, insomuch as they cluster cases (generally using some type of Euclidean distance measure) by searching for a simpler set of reference vectors, with each case in $\{c_i\}$ being positioned near its most similar reference vector. For k-means, the reference vector is a centroid, which represents the average for all the cases in a cluster. For the SOM, the reference vector is an actual point, a neuron, which represents the weighted average of the vector configurations clustering around it.

From here, however, they differ. But, again, it is their differences (in combination with the above similarities) that make them work so well together.

For example, the k-means is useful (and often computed first in an analysis by the SACS Toolkit) because it requires the number of centroids to be identified ahead of time, based largely on theoretical rationale. Such an approach is important for case-based modeling because it requires that the selection of clusters be somehow theoretically driven—researchers should have some sense of what they are looking for, based on a preliminary study and comparison and contrast of the cases $\{c_i\}$ in $\mathcal{N}$. In turn, the SOM functions as an effective method of validating k-means because the set of reference vectors (neurons) it settles upon is unsupervised. If, therefore, the SOM arrives at a solution similar to the k-means, it provides an effective method of corroboration.

The SOM is also useful because it graphs its reference vectors and the cases $\{c_i\}$ surrounding them as neurons on a two-dimensional, topographical map, called the *U-matrix*. On the U-matrix, the reference vectors most like one another are also graphically positioned as nearby neighbors, with the most unlike reference vectors (neurons) being placed the furthest apart. The U-matrix therefore provides a visually intuitive, low-dimensional map of the original high-dimensional (complex) proximity matrix $\mathcal{P}$ and its main clusters. The SOM provides other visualization techniques as well, such as projecting cases $\{c_i\}$ and their clusters into two-dimensional or three-dimensional space, based on the two or three most important eigenvectors (principal components) making up the vector configuration of c_i in D.

Social Practice	Cluster			
	Upper Class Communities	Middle Class Communities	Working Class Communities	Poverty Trap Communities
Household Income	68083	42435	33068	20218
% High School Degree	2.7	14.5	19.2	31.1
Average Years of Life Lost	10.50	12.90	14.37	15.36
	n=1	n=5	n=7	n=7

Fig. 4 K-means cluster solution for *Summit*

7.2.1 Case study

To compress the $\mathcal{P}$ for *Summit*, we followed the above two-step process of k-means and the SOM. We ran our k-means with SPSS and we ran the SOM with the SOM Toolkit. To keep things simple, we did not include urban sprawl or border sharing in our analyses. Starting with k-means, we explored a variety of cluster solutions, from 3 to 7, seeking the solution with the most even distribution of cases. For example, we found that any solution above 4 clusters resulted in three or more clusters with only one community. As shown in Fig. 4, we therefore settled on a four cluster solution: (Cluster 1) upper class communities ($n = 1$); (Cluster 2) middle class communities ($n = 5$); (Cluster 3) working class communities ($n = 7$); and (Cluster 4) poverty trap communities ($n = 7$).

The four centroids in Fig. 4 constitute the average vector configuration for the cases in each cluster. In terms of the first cluster, however, there is only one community, so the centroid is the community.

As a side note, if this paper was only about our case study, before turning to the SOM we would explore why the k-means arrived at its four cluster solution by (a) interpreting the centroids and (b) exploring the arrangement of the cases within each cluster. For example, it is clear from Fig. 4 that poor communities have much worse health outcomes than affluent communities. We will, however, pause to make one important point. Figure 5 plots the distribution of communities within each cluster. Looking at this figure, it seems there is an outlier in cluster 3, along with a degree of variability in clusters 2 and 4. We address this variability next in our review of the SOM.

The last step in compressing the $\mathcal{P}$ for *Summit* was to validate our k-means by running the SOM. Figure 6 provides three maps. On each map the communities were labeled according to their k-means cluster membership to facilitate easy visual corroboration.

The first (Map A) is the u-matrix without topographical features, showing, at the most basic level, how the SOM clustered the 20 communities—we drew circles around each major cluster. The second (Map B) provides important topographical features: the lighter the area (node) between two or more communities, the more conceptually distant they are from each other—the 'c' shaped line we drew shows the highest ridge on the map. The third (Map C) is a variation of the standard u-matrix, as it projects the clusters onto the three most important principal components (social practices) in the database—again, we drew circles around each major cluster, which matches the clusters in Map A.

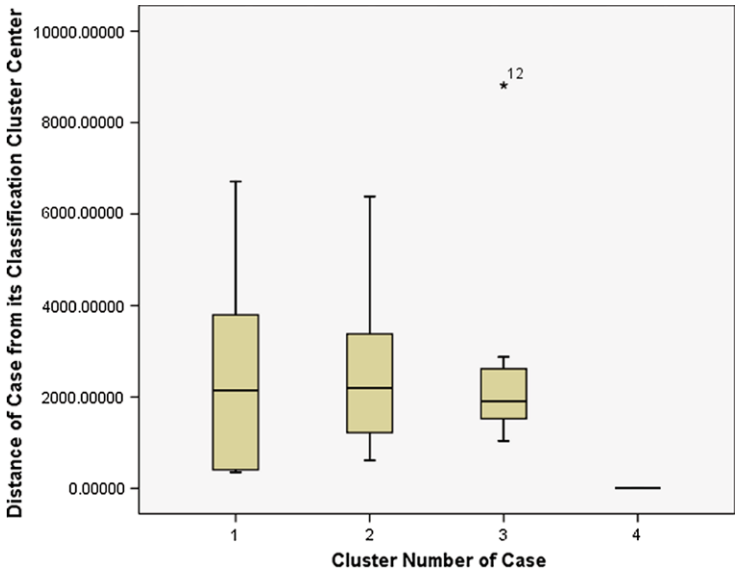


Fig. 5 Within cluster solution for *Summit*

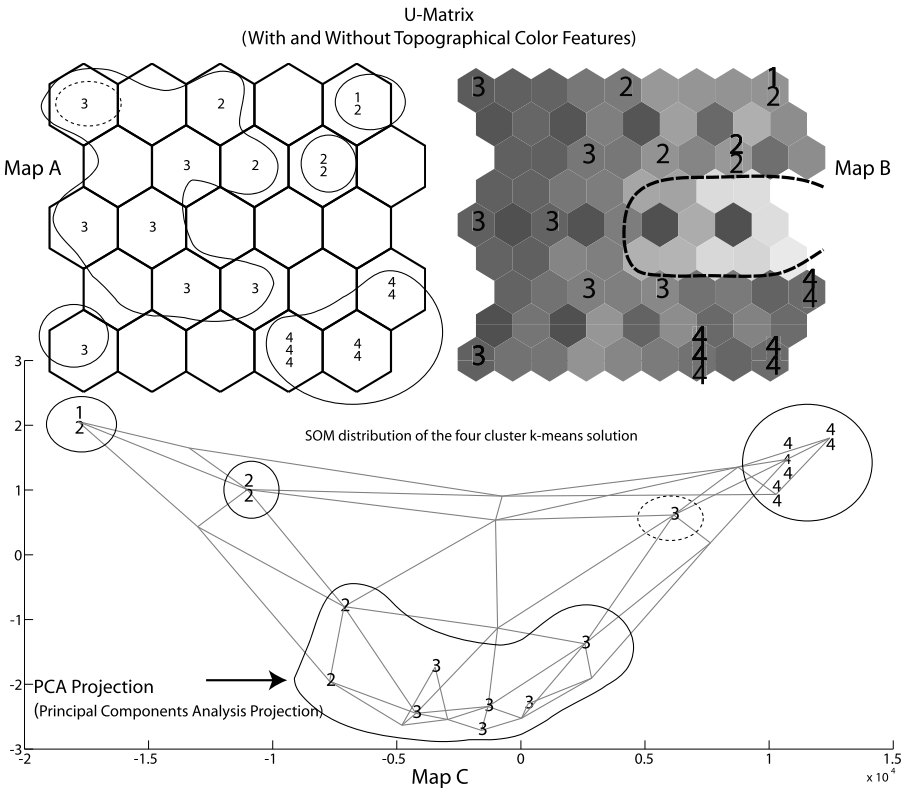


Fig. 6 SOM solution for *Summit*

Looking at the maps, the first thing that stands out is that, on Map A, moving clockwise from the bottom right (where the poorest communities are located) to top left (where the wealthiest communities are located), the SOM distributes the 20 communities similar to the k-means. Map B supports this clockwise placement, as the highest ridge is along the center right.

The second thing that stands out is that the SOM did not group the communities in cluster 2 and 3 as tightly as the k-means did. This difference, however, is primarily a function of intent: as we explained earlier, any k-mean solution beyond 4 generated a significant number of clusters with only 1 community. The SOM seems to corroborate this finding, suggesting that, while the communities in clusters 2 and 3 are similar, there are also meaningful differences within them. If this paper was just about this case study, we could continue to explore these within-cluster differences, seeking to develop a sophisticated understanding of *Summit* in terms of its dominant case-based configurations. However, for the purposes of this paper, we must move on. It is time to generate our set of network maps.

7.3 $\mathcal{N}$ as network map

Network mapping is about compressing $\mathcal{A}$ to model the most important relationships (ties, links, etc.) and interactions that exist amongst the cases $\{c_i\}$ in S, particularly as they relate to the most common vector configurations in S. In terms of compressing the relationships amongst $\mathcal{A}$, the SACS Toolkit primarily uses the methods of the new science of networks and social network analysis. In terms of compressing the interactions amongst $\mathcal{A}$, the SACS Toolkit also uses agent-based modeling. Given the numerous, excellent texts available on these techniques, there is no need to rehearse them here. The only point necessary to make is that the emphasis of this second set of maps is to data-mine further the underlying dynamics and structure of S by (a) exploring the relationships amongst and interactions between the cases in $\mathcal{A}$ and (b) examining how the relationships and interactions within $\mathcal{A}$ relate to the most common vector configurations identified in the compression of $\mathcal{P}$.

7.3.1 Case study

As explained earlier, for our case study we mapped the relationships amongst communities based on border sharing. Figure 7 shows the links amongst the 20 communities in *Summit* based on the borders they share. The reader will note that the communities are labeled according to their cluster membership, not their name. We labeled Fig. 7 this way so we could link our analysis of $\mathcal{A}$ with our compression of $\mathcal{P}$. Looking at Fig. 7, it seems clear that, without even running any statistics, poor communities (clusters 3 and 4) share a significant number of borders with each other and seem to form a cluster on the left side of the map. For example, the community we circled in the upper left is bordered by six other communities, all of them belonging to clusters 3 or 4.

Such clustering suggests (as argued by the current community health science literature) that the vector configurations of poor communities tend to have a high degree of spatial autocorrelation. Again, we could go on to examine this insight statistically

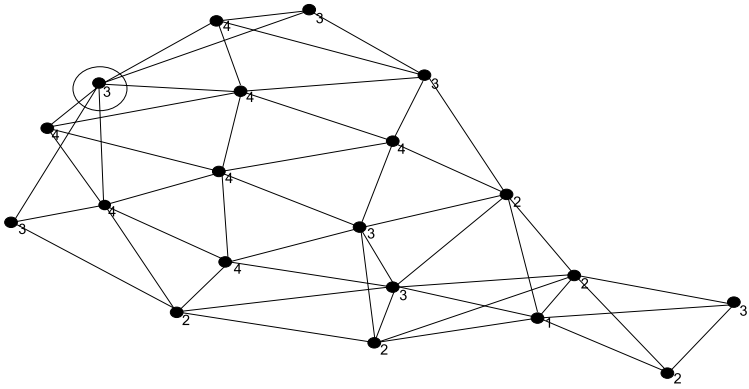


Fig. 7 Network map of border sharing for communities by cluster

and discuss in detail its significance. For example, we could examine why poverty often seems to become geographically clustered. In terms of the current paper, however, the main point is that there is much to be gained in the assemblage of our model of *Summit* by compressing and studying $\mathcal{A}$ and its relationship with $\mathcal{P}$.

We turn, now, to the last set of maps necessary to form our final model.

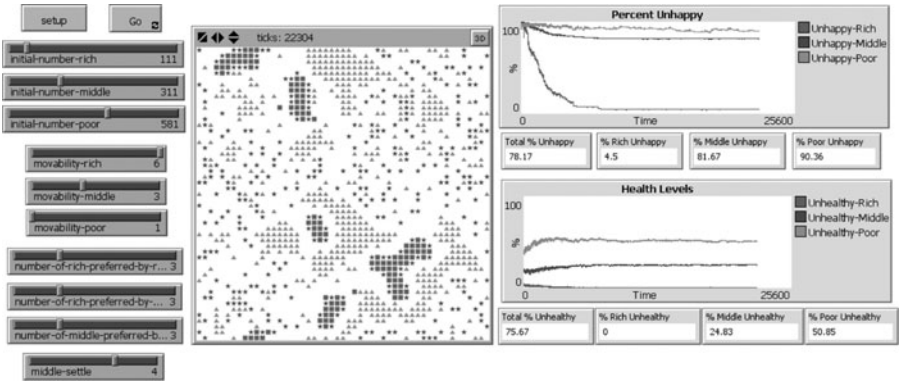
7.4 $\mathcal{N}$ as dimensional map

The third and final set of maps involve compressing $\mathcal{L}$. The goal here is to understand how the coupling of W_s or E_s influence the case-based structure and dynamics of S. This understanding is achieved by exploring three types of relationships.

First, it is necessary to explore how the various k dimensions in $\mathcal{L}$ relate to one another beyond the zero-order level. The focus here is on the search for trends. Again, the methods most useful for conducting such a search are grounded in a well-established and rather extensive literature. And so, a variety of techniques can be employed, from analysis of variance and hierarchical regression to dynamical systems modeling and agent-based simulation.

Second, it is necessary to explore how the various k dimensions in $\mathcal{L}$ (and their relationships with one another) relate to the compression of $\mathcal{A}$. Again, the network methods most useful for conducting such a search are grounded in a well-established and rather extensive literature—in this case, the new science of networks and social network analysis—and therefore do not require review. There is, however, a point to make: the goal of this second set of analyses is to examine how the results achieved from compressing $\mathcal{L}$ relate to the compression of $\mathcal{A}$.

Finally, it is necessary to explore how the various k dimensions in $\mathcal{L}$ (and their relationships with one another) relate to the most common vector configurations identified in the compression of $\mathcal{P}$. Again, there are a variety of ways such an analysis can be conducted. The SACS Toolkit finds two methods particularly useful: k-means and the SOM. K-means is useful because, along with its final cluster matrix, it provides a set of descriptive statistics (i.e., uncorrected, Analysis of Variance, F-scores) on the relative contribution each k dimension in $\{c_i\}$ has on the final cluster solution.



Left-hand side of Dashboard contains all the controls for Summit-Sim; the center window is the 2-dimensional, 51X51 lattice structure in which the world of Summit-Sim takes place; while the right-hand side contains the two output graphs for happiness ratings and healthiness ratings.

Fig. 8 Snapshot of Netlogo Dashboard for Summit-Sim

In turn, the SOM provides a set of principal components maps, each visualizing the relative contribution a particular k dimension of the vector configurations has on the positioning and clustering of $\{c_i\}$ on the final U-matrix.

7.4.1 Case study

For our case study, of the various dimensions to be explored, we focused on urban sprawl and its links with border sharing and years of life lost, as it allowed us to address the three types of relationship explored when compressing $\mathcal{L}$. Specifically, it allowed us to examine the impact sprawl has on the spatial autocorrelation of poverty and health (a.k.a. community vector configurations) in *Summit*, as outlined in Fig. 7 above. Let us explain.

In our in-depth study of Summit County (Castellani et al. 2012), our empirical examination of urban sprawl led us to the following realization: the agent-based (mobility) migration behaviors of agents living in this county changed the composition of several of its poorest and most affluent communities. This change occurred, in large measure, as affluent agents migrated to the more affluent suburban communities located in the northern part of Summit County. The consequence of these residential migration patterns was the spatial segregation of health: poor agents remained confined within or found themselves migrating to poor neighborhoods with poor health outcomes (clusters 3 and 4 in Fig. 6), while more affluent agents migrated to suburban areas with high health outcomes (clusters 1 and 2 in Fig. 6). In turn, the affluent communities became more economically and medically healthy, while the working and poor communities continued to fall behind, creating the socioeconomic-based community clustering we see in Fig. 6. In other words, sprawl in Summit County had a significant impact on the respective vector configurations of its richest and poorest communities and the socio-economic network they form.

To examine the validity of our empirical model, we constructed an agent-based model of *Summit*, called *Summit-Sim*. As shown in Fig. 8, Summit-Sim (which we created using Netlogo) is a rule-based, multi-agent, discrete-event model with a

51×51 lattice structure, upon which a randomly distributed set of upwardly-mobile agents migrate to find their ideal residential location. Living in Summit-Sim are three types of agents, chosen as a rough representation of the current population of Summit County: rich agents (squares, $N = 111$), middle-class agents (triangles, $N = 311$) and poor agents (circles, $N = 581$). While we cannot go into any detail about our agent-based model, it supported our empirical findings: in Summit-Sim, rich agents settled into small neighborhoods with high health and few poor people; middle-class agents clustered into loosely distributed groupings with medium health (like we saw in the SOM's distribution of clusters 2 and 3); and poor agents found themselves stuck in poverty traps with poor health.[2]

As stated several times now, if our case study was the focus of this paper, we would continue to develop our findings, compressing $\mathcal{L}$ in a variety of ways to explore the impact urban sprawl has on community health in *Summit*. However, given the focus of the current paper, we need to stop here.

7.5 $\mathcal{N}$ as discrete model

With the cluster, network and dimensional maps constructed, the next step in the modeling process is to explore further how these maps inform one another, with the goal of arriving at well-developed, albeit simplified, low-dimensional map of S for one discrete moment in time/space.

The construction of a such a model, while highly useful, is not, however, the last step in the modeling process. While researchers need not go any further, the SACS Toolkit was ultimately created to model the structure and dynamics of a complex system as it evolves across time/space. And so the final step in the SACS Toolkit's case-based modeling approach is to address the issue of time/space.

8 Modeling $\mathcal{N}$ across time/space

For the SACS Toolkit, the longitudinal study of S is based on the following set of premises:

(a) Cases c_i are not static; they are dynamic and evolving. Following dynamical systems theory, the SACS Toolkit therefore treats cases c_i as discrete dynamical systems $c_i(j)$, where j denotes the time instant t_j. If cases c_i change across time/space, so too must their vector configurations $c_i = [x_{i1}, \ldots, x_{ik}]$; that is, their measurements on W_s or E_s in D. As such, D is comprised of a series of $c_i(j)$, one for each discrete moment in time/space t_j, on which a set of measurements are taken to construct a particular model of S.

(b) If cases are discrete dynamical systems $c_i(j)$, it therefore follows that, as a final definition, a complex system S is a set of cases $\{c_i\}$, with each case c_i constituting one of its possible ways of practice across time/space, based on the coupling of W_s and E_s.

[2]For those interested in an in-depth review of our model, see Castellani et al. (2012). Also, for those interested in downloading or running Summit-Sim, go to http://cch.ashtabula.kent.edu/summitsim.html.

(c) As discrete dynamical systems, cases $c_i(j)$ have relationships and interactions with one another across time/space. These relationships and interactions can be conceptualized as the edges of a graph, with vertices being the cases $\{c_i\}$.
(d) If the relationships and interactions amongst cases c_i (along with their respective vector configurations) change across time/space, then the field of relations R_s must change as well. R_s (and its three sets of matrices, $\mathcal{P}$, $\mathcal{A}$, $\mathcal{L}$) is therefore comprised of a series of $R_s(j)$, one for each discrete moment in time/space t_j.
(e) If R_s changes across time/space, then it is necessary to construct a network of attracting clusters $\mathcal{N}$ for each of $R_s(j)$. The result of these repeated assemblages of $\mathcal{N}$ is a series of discrete, simplified, low-dimensional maps $\mathcal{N}_j$ of S that have both dynamical and topographical features.
(f) The final goal of the SACS Toolkit is to assemble these $\mathcal{N}_j$ and their respective topographical and dynamic features into a final case-based model of S.

In terms of method, to assemble the $\mathcal{N}_j$ for the evolution of S across time/space, the SACS Toolkit uses the same set of case-based friendly, computationally oriented, methodological techniques previously discussed. The only difference is that, for this final step (as outlined above), the dynamical features of each $\mathcal{N}_j$ (along with its respective topographical features) are addressed. Once these $\mathcal{N}_j$ are assembled into a comprehensive map, the final model is complete.

8.1 Case study

In our actual study of Summit County (Castellani et al. 2012), we had data for two time periods (1990 and 2000), which allowed us to study the longitudinal changes of this county and its 20 communities over time. While we cannot in any way address such details in the current study, it is ultimately the temporal analysis of Summit County that allowed us to understand the dynamics of its community-level health and wellbeing. For example, one of the key insights we arrived at was that, despite efforts to improve the health of poor communities in *Summit*, sprawl (amongst other key factors, such as job growth, etc.) seems to have trapped these communities in some type of poverty trajectory. But, this is only as far as we can go in the current paper. We must stop now and turn to our conclusions.

9 Conclusion

As this paper has hopefully shown, the SACS Toolkit offers researchers a new method for quantitatively modeling complex social systems, based on a case-based, computational approach to data analysis. Developed as a variation on Byrne (2009), the SACS Toolkit models a complex system as a set of n-dimensional vectors (cases), which it compares and contrasts, and then condenses and clusters to create a low-dimensional model (map) of a complex system's structure and dynamics over time/space. The assembled nature of the SACS Toolkit is its primary strength. While it is grounded in a defined mathematical framework, it is simultaneously open-ended and therefore adaptable and amenable, allowing researchers to employ and bring together a wide

variety of computational, mathematical and statistical methods—a very important asset in the methodologically disjointed field of complexity science today. Researchers can even develop and modify the SACS Toolkit for their own purposes. The other strength of the SACS Toolkit, which makes it a very effective technique for modeling large databases, is its ability to compress complex data matrices while preserving the most important aspects of a complex system's structure and dynamics over time—a very important ability in the new era of data overload. The next step in the process of developing the SACS Toolkit is for researchers to explore some of its more technical aspects, with the most important being how to model complex systems across time/space.

References

Ahearn L (2001) Language and agency. Annu Rev Anthropol 30:109–137
Bourdieu P (1990) The logic of practice. Polity Press, Cambridge
Byrne D (2009) Using cluster analysis, qualitative comparative analysis and NVivo inRelation to the establishment of causal configurations with pre-existing large-N datasets: machining hermeneutics. In: Byrne, Ragin's (eds) The sage handbook of case-based methods. Sage Press, Thousand Oaks, pp 260–268
Castellani B, Rajaram R, Buckwalter JG, Ball M, Hafferty F (2012) Place and health as complex systems: a case study and empirical test. Proc Center Complex Health 1(1):1–35. Kent State University at Ashtabula
Castellani B, Hafferty F (2009) Sociology and complexity science: a new field of inquiry. Springer, Berlin
Castellani B, Hafferty F (2010) The increasing complexities of professionalism. Acad Med 85:288–301
Cummins S, Curtis S, Diez-Roux AV, Macintyre S (2007) Understanding and representing place in health research: a relational approach. Soc Sci Med 65:1825–1838
Dreyfus H, Rabinow P (1983) Michel Foucault: beyond structuralism and hermeneutics, 2nd edn. Chicago University Press, Chicago
Foucault M (1980) In: Gordon C (ed) Power/knowledge: selected interviews and other writings 1972–1977. Pantheon Books, New York
Giddens A (1980) The constitution of society: outline of the theory of structuration. University of California Press, Berkeley
Jain A (2009) Data clustering: 50 years beyond k-means. Pattern Recognit Lett
Janasik N, Hankela T, Bruun J (2008) Text mining in qualitative research: application of an unsupervised learning method. Organ Res Methods 20(5):1–26
King A (2004) The structure of social theory. Routledge, New York
Newman M (2010) Networks: an introduction. Oxford University Press, Oxford
Ragin C (2008) Redesigning social inquiry: fuzzy sets and beyond. University of Chicago Press, Chicago
Rihoux B, Ragin C (2009) Configurational comparative methods: qualitative comparative ANalsysi (QCA) and related techniques. Applied social research methods series, vol 51. Sage Press, Thousand Oaks
Sperandio M, Coelho J (2004) K-means and som, a concurrent validation scheme for data mining. In: Intelligent engineering systems through artificial neural network, vol 14. ASME Press, New York, pp 289–294
Uprichard E (2009) Introducing cluster analysis: what it can teach us about the case. Sage Press, Thousand Oaks, pp 132–147

Brian Castellani is Director of the Center for Complexity in Health, Robert S. Morrison Health and Science Building, Kent State University at Ashtabula. His research focuses on developing the SACS Toolkit and applying it to the study of health and health care issues.

Rajeev Rajaram Department of Mathematics, Kent State University at Ashtabula, is a member of the Center for Complexity in Health. His research focuses on stability and control theory of differential equations and the development of the SACS Toolkit.