RESEARCH ARTICLE

Cases, Clusters, Densities: Modeling the Nonlinear Dynamics of Complex Health Trajectories

BRIAN CASTELLANI,[1] RAJEEV RAJARAM,[2] JANE GUNN,[3] AND FRANCES GRIFFITHS[4]

[1]Department of Sociology; [2]Department of Mathematics, Kent State University, Ashtabula, Ohio 44004; [3]Melbourne Medical School, University of Melbourne, Carlton, Victoria, 3053, Australia; and [4]Warwick Medical School, University of Warwick, Coventry, CV4 7AL, UK

Received 20 April 2015; accepted 31 August 2015

In the health informatics era, modeling longitudinal data remains problematic. The issue is method: health data are highly nonlinear and dynamic, multilevel and multidimensional, comprised of multiple major/minor trends, and causally complex—making curve fitting, modeling, and prediction difficult. The current study is fourth in a series exploring a case-based density (CBD) approach for modeling complex trajectories, which has the following advantages: it can (1) convert databases into sets of cases (k dimensional row vectors; i.e., rows containing k elements); (2) compute the trajectory (velocity vector) for each case based on (3) a set of bio-social variables called traces; (4) construct a theoretical map to explain these traces; (5) use vector quantization (i.e., k-means, topographical neural nets) to longitudinally cluster case trajectories into major/minor trends; (6) employ genetic algorithms and ordinary differential equations to create a microscopic (vector field) model (the inverse problem) of these trajectories; (7) look for complex steady-state behaviors (e.g., spiraling sources, etc) in the microscopic model; (8) draw from thermodynamics, synergetics and transport theory to translate the vector field (microscopic model) into the linear movement of macroscopic densities; (9) use the macroscopic model to simulate known and novel case-based scenarios (the forward problem); and (10) construct multiple accounts of the data by linking the theoretical map and k dimensional profile with the macroscopic, microscopic and cluster models. Given the utility of this approach, our purpose here is to organize our method (as applied to recent research) so it can be employed by others. © *2015 Wiley Periodicals, Inc. Complexity 000: 00–00, 2015*

Key Words: longitudinal data; case-based modeling; nonlinear dynamics; complex health trajectories; differential equations; vector quantization

Correspondence to: B. Castellani, Department of Sociology, Kent State University, Ashtabula, Ohio 44004. E-mail: bcastel3@kent.edu

1. INTRODUCTION

Modeling the nonlinear dynamics of complex health trajectories across time presents a number of serious challenges for scientific inquiry [1–5]. The

 Vol. 00 No. 00
DOI 10.1002/cplx.21728
Published online 00 Month 2015 in Wiley Online Library (wileyonlinelibrary.com)

challenge comes in the form of method (both in terms of the complexities of data and the limitations of conventional techniques).

In terms of data, the challenge is that complex trajectories, be they cohort or longitudinal data: (1) seldom follow a singular common trend; instead, (2) they self-organize into multiple major and minor trends; which, (3) when modeled microscopically, are highly dynamic and complex—often taking the form of a variety of complex behaviors—making curve fitting, prediction and control (for example, health management) very difficult; furthermore (4) these continuous trends are often a function of different measurements (k dimensional vectors) on some profile of biomedical-psycho-social factors and (5) the complex set of qualitative interactions and relationships amongst these variables [1–3].

Our last point on data leads to the challenge of technique: (1) while medicine and health are ultimately about the case – k dimensional vector profiles – health researchers tend to ignore these complex profiles and the set of qualitative interactions of which they are comprised; (2) focusing, instead, on what they deem to be the most salient (and relatively independent) handful of variables relevant to some outcome of concern; (3) which they model by controlling for the remaining profile of variables; (4) furthermore, they study these few factors using some form of linear (and often discrete) modeling/statistics; (5) typically in the search for the most common one or two aggregate trends [1–3].

As a result of this approach, there is a major disconnect between health data and health research, making it difficult for scientists to do such things as (1) model the aggregate nonlinear dynamics and complex trajectories of cases or their densities in continuous time; (2) detect the presence of multiple trends (i.e., major and minor) across time; (3) identify and map complex steady-state behaviors (i.e., transient sinks, spiraling sources, periodic orbits); (4) explore and predict the motion of different health trajectories and time instances; or (5) link these different trends to the complex k dimensional vectors/profiles upon which they are based, so that (6) they can construct a multilevel theoretical model of their topic of study [4–7].

While the above disconnect between health data and technique is problematic, researchers are beginning to 'turn' to the complexity sciences and computational modeling and related areas of inquiry—genetic algorithms, dynamical systems theory, network analysis, differential equations, control theory, etc—for possible solutions [8]. In regards to this 'complexity turn,' we seek to demonstrate the utility of case-based complexity for modeling complex health trajectories [9].

Case-based complexity combines case-comparative method with the various theoretical and methodological tools of the computational and complexity sciences to advance the modeling of complex (social and health) systems, which it does by treating complex systems as sets of cases (i.e., k dimensional vectors/profiles) [4–7]. The platform we created for this approach is called the SACS Toolkit [4–7].

The SACS Toolkit is a case based, computationally grounded, mixed methods framework for modeling complex systems. One of its key strengths—called a case-based density approach—is its capacity to model the nonlinear dynamics of complex trajectories, particularly in the form of cohort or longitudinal data [4,5]. To do so, it employs a novel combination of case-comparative method in conjunction with vector quantization, genetic algorithms, ordinary differential equations (ODE), Haken's synergetics, the inverse-forward problem, and nonequilibrium statistical mechanics, specifically transport theory and the continuity (advection) partial differential equation (PDE). The result is a ten-step, multilevel procedure for transforming the nonlinear dynamics of complex trajectories into cases, clusters, and densities [4–7].

The current study is fourth in a series. [4,5,7] The purpose of the first three papers was to provide a mathematical outline of our approach [7] and to work on several key steps, including (1) a technique for fitting an ODE directly to data and (2) a procedure for using the vector field thus obtained to simulate the evolution of the distribution of cases (as densities) across time using the advection PDE [4,5]. Still, the following remains to be done. We have yet to:

1. assemble our ten steps into a formal outline for others to employ;
2. highlight how the various and multiple outputs of our ten steps go together to create our multi-level model; and
3. demonstrate the utility of our approach in application to several of our recent studies.

Hence, we come to the purpose of the current study: we seek to formalize our case-based density approach, as employed through the SACS Toolkit, in application to several of our recent health studies, including a study on allostatic load [10], public health [1], and international health [5], along with a forthcoming study on depression and wellbeing [11]. While the last three studies are all longitudinal, the first is discrete; nonetheless, we will refer to it here, as it was crucial to developing several of our steps.

2. MODELING HEALTH TRAJECTORIES: CASES, CLUSTERS AND DENSITIES

Case-based density modeling, as employed through the SACS Toolkit, is a *ten*-step, multilevel procedure for studying the nonlinear dynamics of complex trajectories, the process of which can be summarized as follows:

DOI 10.1002/cplx

2.1. Steps 1 Through 4: Cases, Traces, and Profiles

The purpose of the first four steps is to construct a case-based portrayal of the topic (complex system) of study by: (1) rethinking the database from a case-based (as opposed to a variable-based) perspective; (2) computing the trajectory (velocity vector) for each case, based on (3) an identified set of variables (which, in case-based terms, we call traces – which we will explain below); and (4) constructing a working theoretical model to explain the trajectory of these traces.

2.1.1. Step 1

The first step is an epistemological one: it requires a cognitive shift from a variable-based to a case-based view of the topic (complex system *S*) of study. In doing so, we follow Byrne [12] and Ragin [13,14] and the notion of casing, which allows us to (a) treat our topic of study as a complex system and (b) in turn, vary our notion of what the 'case' is, depending upon different empirical concerns.

In terms of the first point, according to case-based complexity, cases are complex profiles comprised of a set of inter-dependent variables, which are contextually dependent, nonlinear, dynamic, evolving, self-organizing, emergent, etc. in short, cases have the same characteristics as a complex system. Theoretically speaking, then, cases can be treated and modeled as complex system. [8,9].

It is, nonetheless, important to point out that, while cases can be treated as complex systems, not every and any set (profile) of variables can be treated as such. As Byrne [12] and Ragin [13,14] make clear, a case-based approach requires theoretical rationale, grounded in empirical support. (For more, see Byrne [12] and Ragin [13,14])

In terms of the second point, however, the cases for any given study can vary significantly. For example, a case may be a theoretical construct, a commonly recognized empirical unit, or some combination thereof. Cases can also form part of or contribute to a complex system *S* of study. Or, cases may be considered as nested within a complex system *S*. We will call these types of 'nested' cases 'second-order' for now. In such instances, there is interpenetration of the complex system *S* and these second-order cases. [15] Finally, and yet again, a complex system *S* might be considered as nested within some set of second-order cases. In short, numerous possibilities exist.

Nonetheless, to make our approach clear, for us, in the initial stages of analysis, we primarily treat cases as complex configurations, similar to Ragin. [13,14]. Furthermore, we often do see cases and complex systems as nested in one another. And, following Byrne, in the final stages of analysis, when we get to constructing our multiple narratives, we treat cases in more 'storied' terms. But, overall, we are primarily grounded in the idea of complex systems as comprised of sets of cases, which emerge out of the configuration of k dimensional row vectors – which we will discuss in a moment.

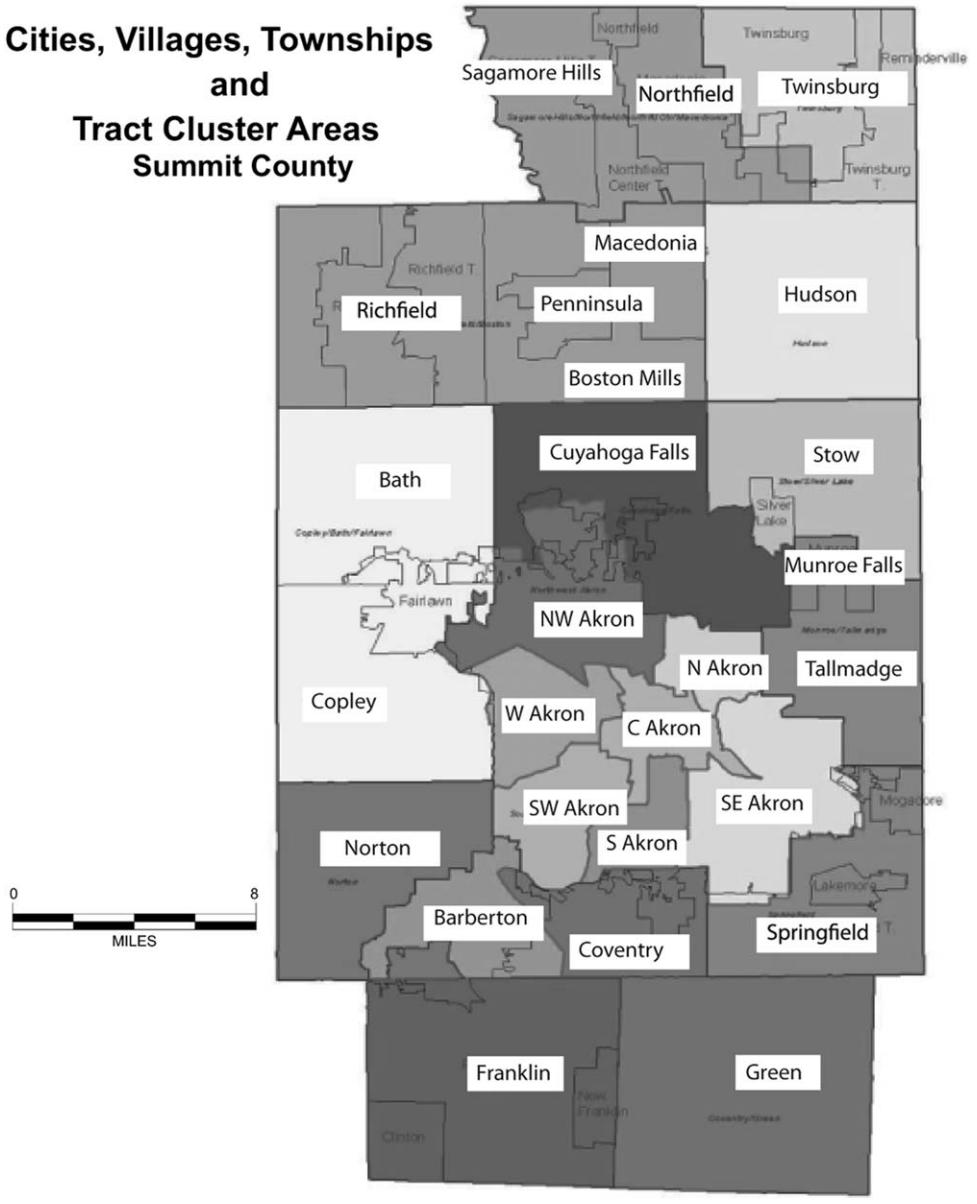
By way of example, here are two illustrations where we reconceptualized a topic of study in case-based, complex systems terms. In a recent public health study, we conducted on a Midwestern county in the United States [1], our case was a County, which we conceptualized (using census data) as a set of 20 communities (smaller cases). As such, as shown in Figure 1, for our study we moved back and forth between Summit County (our primary case) and its twenty major communities (our more specific cases). Here, the nesting of the more specific cases within our primary care was relatively clear, and the casing was both empirical and conceptual. Furthermore, following Cilliers, the boundary for our primary case was "simultaneously a function of the activity of the system itself, and a product of the strategy of description involved" [15], as well as a coupling between our primary case and its nested hierarchy of cases, and their combined evolution across time/space. [12]

As a second example, in a recent study, we conducted on the negative impact of stress, we treated allostatic load as our primary case [10]. (This is a less conventional use of casing than the example of county and community, albeit medicine is, in practice, based on the idea of cases and casing.) We did this by conceptualizing it as a complex clinical construct, comprised of a large number of interdependent biological subsystems, which are represented by an even larger number of interconnected biomarkers. However, when it came to our database, allostatic load became more of an abstraction, as we did not seek to build a single model of this clinical system. Instead, we sought to treat each of our $N = 1151$ cases (people) as an individualized model. In other words, each case in our study constituted one possible way that allostatic load manifests itself in people's lives; one possible trajectory in the larger state-space of all possible trajectories, based on the unique way allostatic load self-assembles itself for each case. Furthermore, there is clearly coupling taking place between allostatic load (the primary case), and the individuals in the study living their life (the second-order cases.)

As these two examples illustrate, it is this unique, case-based approach that distinguishes our method from the 'single-model, variable-based approach' common to many methods in statistics and mathematical modeling. However, in turn, it makes our approach similar to more recent developments in data mining and data sciences, as well as such techniques as agent-based modeling, Markov models, and latent growth and mixed-methods modeling—which, by the way can be (and have been) employed in conjunction with our technique. Similar to these

DOI 10.1002/cplx

FIGURE 1



Map of twenty communities in Summit county.

approaches, we begin with the assumption that any complex system of study requires multiple and different (albeit interconnected) case-based models, as there is no one trajectory taken by the system's cases. Instead, cases follow and cluster together along multiple major and minor trends, which we will discuss below. And, following Ragin, [13,14] we assume that systems and cases are often nested, requiring multi-level and hierarchical modeling. With this epistemological shift in thinking established, next the database requires further reconceptualization.

From a case-based complexity perspective, each row in a study's database D becomes a complex case c_i, where each c_i is a k dimensional row vector $c_i = [x_{i1}, \ldots, x_{ik}]$ and where each x_{ij} represents a measurement on the profile of longitudinal variables (traces) for D – what case-based researchers call the case profile. For example, in our public health study (Figure 2) we treated each of the 20 cases (communities) as a set of measurements on an in-depth profile (k dimensional row vector) of contextual, compositional and health factors. Furthermore, these variables

DOI 10.1002/cplx

FIGURE 2

Compositional Factors	• Population 65 years of age of older[1] • % White Population[1] (Defined as number of persons identifying themselves as "White" in response to the 1990 US Census or "White Alone" in response to the 2000 US Census) • % African-American Population[1] (Defined as the number of persons identifying themselves as "Black or African-American" in response to the 1990 US Census or "Black or African-American Alone" in response to the 2000 US Census) • Median Household Income[1]
Contextual Factors	• Overall Poverty[1] (Defined as the number of persons living "below the poverty level" as defined by the U.S. Census) • Public Assistance[1] (Defined as the number of households receive public assistance as defined by the U.S. Census) • Persons 25+ Years with High School Diploma[1] • Net Job Growth[3] (Defined as the number of jobs in 2000 minus the number of jobs in 1990. • Unemployment Rate[1] (Defined as unemployed civilian labor force) • Housing affordability[1] (Defined as the percentage of households where mortgage/rent is greater than 30% of the household income) • No Health Care Coverage[4] (An estimate of the number of individuals with no health care coverage based upon a statewide survey (Behavior Risk Factor Surveillance Survey – Centers for Disease Control and Prevention)
Health Outcomes	• No First Trimester Prenatal Care[4] (Defined as the number of births occurring to mothers from 1995 to and including 1998 for which no prenatal care was received during the first three months of the pregnancy) • Teen Birth Rate[4] (Defined as the number of births occurring between 1995-1998 to mothers 15 to and including 17 years of age) • Childhood Immunization Rate[5] (Defined as the percentage of children with a complete immunization series 4:3:1 by their second birthday based on the kindergarten retrospective study) • Child Abuse/Neglect[6] (Defined as the number of referrals resulting in assessment per 1,000 childre under 18 years of age) • Elder Abuse/Neglect[7] (Defined as the number of referrals received by the Department of Jobs and Family Services for abuse, exploitation, or neglect) • Years of Potential Life Lost per Death[5] (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 due to all causes divided by the number of deaths due to all causes within the census tract cluster borders where those borders are defined by United States Census Bureau census tracts)

Data Sources: (1) United States Census Bureau 1990 and 2000 Decennial Censuses; (2) Ohio Department of Education; (3) NODIS; (4) Akron City Health Department, Office of Epidemiology; (5) Ohio Department of Health; (6) Children's Services Board; (7) Summit County Department of Jobs and Family Service.

Variables used to assemble case-based profile for 20 communities.

(traces) were measured at two discrete time points (2000 and 2010). In turn, in a recent international health study, we examined the longitudinal relationship between per-capita GDP and human longevity rates (our two profile variables) for 156 countries (each a complex dynamical case) over the course of 63 years. Data for this model came from the widely used Gapminder dataset http://www.gapminder.org/.

The temporal nature of these two examples take us to our next point: cases c_i are not static; instead, they are dynamic and evolving. As such, in terms of cohort and longitudinal databases D, each case c_i in D is, ultimately, a complex dynamical system $c_i(j)$, where j denotes the time instant t_j. In turn, if the trajectories of cases c_i change across time/space, so too must their vector configurations $c_i=[x_{i1},\ldots,x_{ik}]$. As such, in terms of cohort and longitudinal studies, D is comprised of a series of $c_i(j)$, one for each moment in time/space t_j (discrete or continuous), on which a set of measurements are taken to construct a particular model of the complex system of study S.

Two examples: First, in a new study we are conducting on depression and wellbeing, we examined seven years worth of trajectory data [11]. Data for this study ($N=259$ cases) came from a subsample of the Diamond Prospective Longitudinal Cohort Study, one of the largest primary care depression cohort studies worldwide. [2,3] Second, in our international health study (as mentioned above) we examined the longitudinal relationship between per-capita GDP and human longevity rates for 156 countries over the course of 63 years. In fact, Figure 3 shows a microscopic model of the trajectories of these countries across time (shown in blue) along with the model we 'fitted' to the data (shown in green). The X-axis represents GDP; and the Y-axis represents life expectancy.

2.1.2. Steps 2 and 3

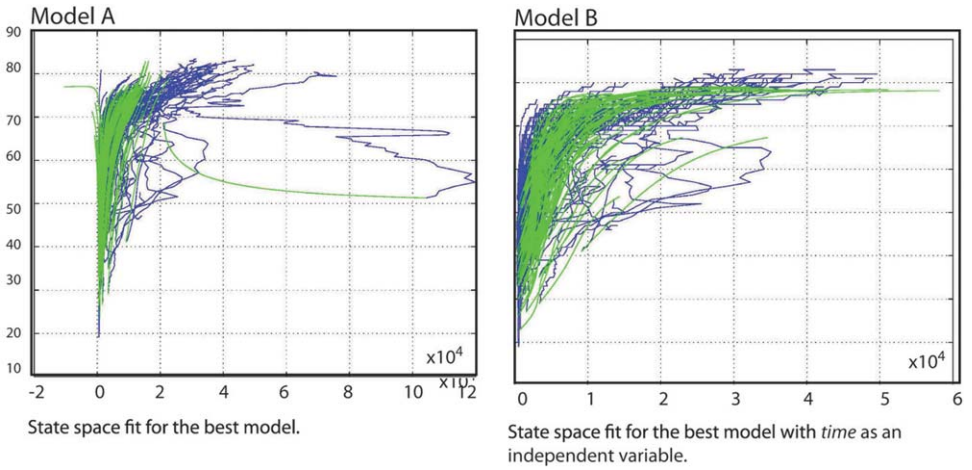
With the database reconfigured into a case-based framework, the next two steps are to identify and model the key traces of the system.

The challenge with modeling cohort and longitudinal data is that, given some complex profile of study, the resulting vector configuration $c_i=[x_{i1},\ldots,x_{ik}]$ and corresponding vector field are k dimensional and therefore too complex or dynamic to be accurately modeled [8,16]. As a result, even with a working map in hand, one rarely has direct access to the actual state of the system studied. Instead, one studies the system's state in a modified form.

Drawing upon the work of Byrne and Callaghan [8], we refer to this modified form of the system as they do, defining it as its trace. (For those interested in a detailed

DOI 10.1002/cplx

FIGURE 3



State space fit for the best model.

State space fit for the best model with *time* as an independent variable.

Case trajectories for 156 countries with model fitted to data.

account and defense of this concept, see [8].) In other words, while the ultimate modeling goal of case-base complexity is idiographic analysis, one never fully models the complete complexity of a case or set of cases; instead, one only studies their traces, albeit from a case-based, complex systems perspective.

For example, in a typical health study, one often only has access to the measurements that were made for each case; and these measurements (temperature, blood pressure etc.,) are simply observables exhibited by the human body. The true state (by contrast) is the actual blood flow through the arteries of the body or the ability of the body to cope with impending disease.

By way of another example, we can turn to control theory. In control theory, the ability of one's identified measures to accurately indicate the true state of the system is (roughly speaking) defined as the observability of the system under study. Highly observable systems, for example, allow for measurements that are very good indications of the true state; by contrast, systems with low observability have measurements that are not able to truly capture the state of the system through the external measurements. The trace variables (which are often selected after much consultation with subject matter experts) are, in general, those that improve the observability of complex system—often as first quantitative indicators. [17] For a precise definition of observability see [18]

Theoretically speaking, for [8] a trace can be anything measurable that is directly/indirectly influenced by the actual state of a system. In practice, however, given our complex-systems perspective, we find it useful to begin with the output/dependent variables in a case-based profile. Two reasons:

First, the output is typically what researchers are trying to understand, model, manage or control. For example, in our study of public health, we were primarily interested in community-level health outcomes across time (Figure 2); and, in our study of international health, the focus was on human longevity (across time) in each of our 156 countries (Figure 3). Nonetheless, once these output traces are explored, one continues onward to increase the complexity of the study by exploring their intersection with other key traces—which takes us to the next point.

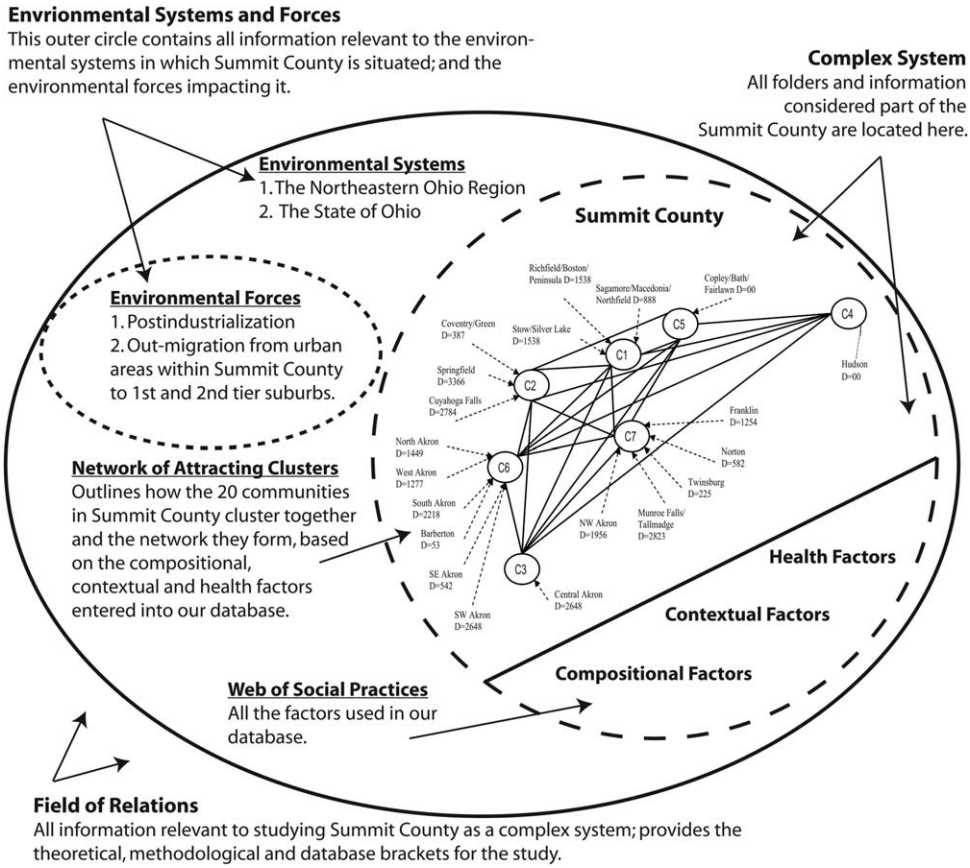
Second, starting with the dependent traces provides a useful way to avoid becoming bogged down in the multiple traces and their interactions within the case-based profiles. As we will discuss later, the purpose of Step 10 is to explore how the traces (as outcomes, outputs, dependent variables) link to, evolve or change in relationship to other key biological, psychological, social or ecological traces. For example, in our public health study, we examined our community-level health outcomes in relation to the compositional and contextual factors shown in Figure 2; and, in our international health study, we examined human longevity rates in each country in relation to that country's per-capita GDP.

With the traces identified, the next thing is to use them to compute the velocity vector for each case. Later, in Steps 5 and 6, we detail the process of computing the velocity vector field. Here we want to provide our rationale for why velocity vectors are so important to our approach.

A key feature of our approach, mathematically speaking, is our link between an 'algebraic-based' definition of cases as k dimensional vectors and a 'calculus-based' definition of vectors as quantities with direction and magnitude. We make this link for several reasons: it allows us to

DOI 10.1002/cplx

FIGURE 4



Theoretical map of twenty communities and trace variables in Summit county.

(1) treat cases (when possible) as continuous trajectories (traces); (2) compute these continuous trajectories (traces) as a function of change between time-stamps (something statistics struggles, at best, to do); (3) examine these changes as a function of discrete or continuous measures on the k dimensional vector for each case (which brings in our theoretical model); (4) employ ODEs to explore the rate of change or velocity of cases (first-order ODE), as well as acceleration (second-order ODE) if needed; and (5) link steps 1 through 4 with the modeling processes in steps 5 through 10. As such, computing the velocity vector for each case functions as *the* main methodological link upon which our approach is based. (As a side note, discrete or hybrid trajectories can be handled by the use of nonlinear difference equations instead of differential equations. We have not tried this approach yet, but presume that it will be an easy implementation.)

2.1.3. Step 4

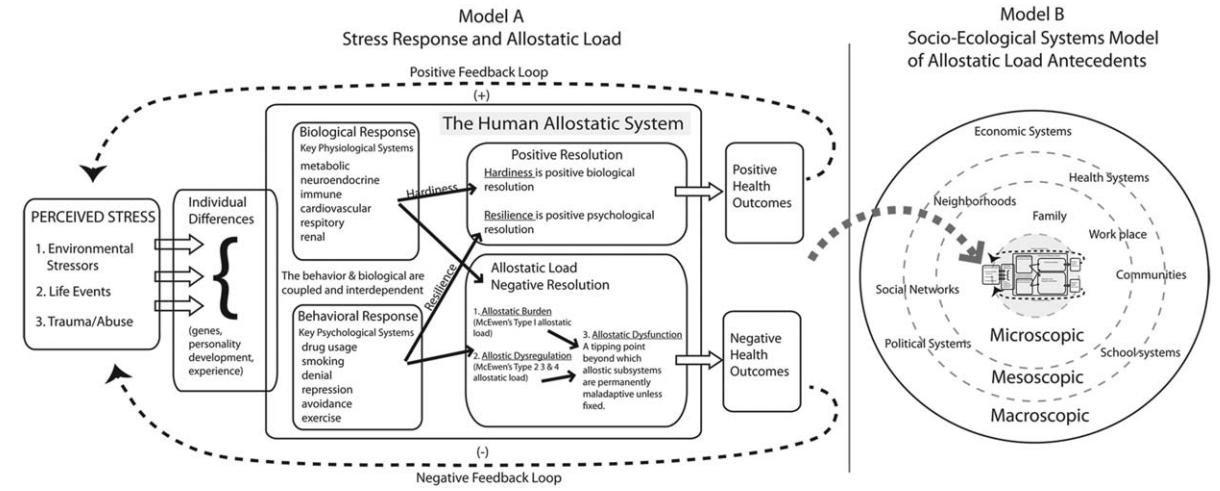
With the velocity vectors computed, the next step is to construct a working theory (map) of the topic of study. The purpose of this map is to theorize, albeit tentatively, how the factors in the case-based profile—as a complex system of interacting variables—go together in relation to the set of outcome(s) being observed, which are treated as trajectories (traces) across time/space.

For example, to arrive at our conceptualization of community health as a complex system we needed a theoretical map. The result was Figure 4. The utility of this map is that it gave us an idea of what traces to explore and how to think about their complex inter-relationships, particularly in relation to our community health outcomes (which we will discuss later). For example, looking at the map, one sees the key factors outlined in Figure 2 (compositional and contextual) as well as some of the key environmental forces we explored in our study; also one sees the three types of maps we constructed for our study, including (as we will discuss in Step 10) a social network analysis of the relationships amongst our 20 communities.

Another example of the importance of a theoretical model is our study on allostatic load. A shown in Figure 5,

DOI 10.1002/cplx

FIGURE 5



Theoretical map of allostatic load as complex clinical construct.

the utility of this map was that it gave us an idea of what traces to explore and how to think about their complex interrelationships, particularly in relation to various health risk outcomes (which we will discuss later). Using this map, we worked with context experts to settle on 20 key biomarkers—each constituting one of the key variables (traces) in our k dimensional profile. We then factor analyzed these 20 biomarkers to construct a seven-factor solution, as shown in Table 1. In turn, these seven factors became our seven main trace variables.

2.2. Step 5: Major and Minor Clusters and Trends

With the theoretical map constructed, the fifth step is to identify the major and minor trends in the data, based on the traces initially chosen for study. (Note: for us, 'major' refers to clusters with high membership and 'minor' refers to relatively lower memberships.) This step is done using k-means cluster analysis and the topographical neural net known as the self-organizing map [19–22]. Our approach is unique in three important ways:

2.2.1. Knowledge-Free, Unsupervised Clustering

First, it is unique in that we take an unsupervized approach to trend identification. Based on the multiplicity of possible trajectories generated by the theoretical model from Step 4, we do not make any reductive or retroductive assumptions about the number of cluster trajectories or possible major and minor trends in the data. Instead, we strive to identify these trends first and then model them separately for each trace, thereby allowing for the creation of multiple models for the same system.

Unsupervised in this context, therefore, has a very precise meaning. Following [19–22] it means that, at this stage in the modeling process, the data are examined as longitudinal trajectories without any context, then converted to z-scores to remove bias, and then clustered using known methods. It is only after identifying the major and minor trends that these case trajectories are examined to determine their empirical veracity and theoretical utility. It is in this sense that we use the term knowledge-free. By contrast, identifying the trace variables is not knowledge-free, because everything starts with the context of data, as well as consultation with subject matter experts and theory—which is where we differ significantly from the 'anti-theory' trend in big data. [23,24] In other words, while identifying the traces requires knowledge, it is not the case with the major and minor longitudinal trends, which are identified without any bias towards what we might expect the trends to be. Said another way, we let our multiple clustering methods corroborate themselves mathematically—which is the goal of such machine driven techniques as k-means, SOM, latent growth modeling, etc—and then investigate those major and minor trends to assign contextual meaning.

For example, in our study of depression and physical wellbeing, we identified eighteen different cluster trajectories. And, in our public health study, we arrived at a seven-cluster solution. Finally, in our allostatic load study, we identified nine clusters, which are shown in Table 2.

Despite the differences in outcome, in all the studies the goal was the same: to allow the data, through the key traces identified, to interact with, speak to, temper,

DOI 10.1002/cplx

TABLE 1

The Seven Factors (Traces) for Allostatic Load

Biomarkers	Factors/Components[a]						
	Blood Pressure	Metabolic Syndrome	Cholesterol	Proinflammatory Elements	Stress Hormones	Blood Sugars	Stress Antagonists
Systolic BP[b]	**0.880**	0.158	0.060	0.132	0.054	0.130	−0.106
Diastolic BP[b]	**0.883**	0.181	0.120	−0.052	0.141	0.020	0.220
Waist to hip ratio	0.305	**0.700**	−0.090	0.113	0.150	0.308	0.294
HDL[c]	−0.096	**−0.829**	0.103	−0.084	−.0191	−0.129	−0.122
Insulin	0.082	**0.677**	0.030	0.379	0.025	0.411	−0.007
Triglycerides	0.164	**0.786**	0.297	0.113	0.039	0.235	−0.093
Total cholesterol	0.099	−0.005	**0.980**	0.021	−0.033	0.011	−0.011
LDL[d]	0.098	.095	**0.935**	0.021	0.040	−0.077	0.093
IL6[e]	0.030	0.271	−0.141	**0.786**	0.000	0.169	−0.257
Fibrinogen	0.001	−0.009	0.092	**0.804**	−0.037	0.148	−0.096
C Reactive Proteins	0.071	0.249	0.100	**0.816**	0.033	0.185	−0.259
Cortisol	0.094	−0.046	−0.008	−0.119	**0.613**	−0.093	0.264
Norepinephrine	0.124	0.237	0.006	0.124	**0.889**	0.075	−0.001
Epinephrine	0.112	0.077	−0.028	−0.085	**0.855**	−0.016	0.178
Dopamine	0.044	0.190	0.000	0.020	**0.888**	−0.006	0.124
Hemoglobin A1c	0.036	0.208	−0.059	0.238	−0.018	**0.887**	−0.163
Glucose	0.115	0.355	−0.015	0.130	0.006	**0.895**	−0.015
DHEAS[f]	−0.005	0.127	0.110	−0.098	0.226	−0.005	**0.729**
Peak Flow	0.208	0.307	−0.089	−0.286	0.111	−0.004	**0.629**
IGF−1[g]	0.031	−0.081	0.020	−0.190	0.026	−0.162	**0.719**

[a]The allostatic load factor structure was obtained using a principal components analysis with promax solution. Biomarkers were retained for the factor on which they loaded the highest, with a minimum loading of 0.613.

[b]Blood pressure.

[c]High density lipoprotein.

[d]Low density lipoprotein.

[e]Interleukin 6.

[f]Dehydroepiandrosterone sulfate.

[g]Insulin-like growth factor.

impact, disagree with, modify, or corroborate the theoretical model. The model that proved the best fit, based on its corroboration with our theoretical map, is the one used.

That is not, however, where the search for trends needs to stop. We can go on to identify subtrends (and even sub-sub-trends) for any particular cluster (as long as the data support this), thereby potentially giving us a hierarchy of models that range from a single model for the entire database of cases (which may not achieve much more than conventional analytic approaches) all the way down to a model for each individual case (which is unlikely to provide novel or useful insights).

It is because of our knowledge-free approach to clustering that our method is a data-driven special case of the inverse-forward problem: we start with data analysis to identify trends; then we move to and develop the theoretical model to organize the causal mechanisms for the trends; then we go back to the data to identify sub-trends if need be and also model the mechanisms identified; and then back to the theoretical model. As the data grows in size with the addition of cases or time-stamps, we can repeat the process, hence the method scales as well.

2.2.2. Longitudinal Clustering

Second, it is unique in that we use vector quantization to engage in longitudinal clustering. We need to emphasize that Step 5 involves clustering case trajectories; not static profiles, as is done in traditional clustering. To cluster cases longitudinally, we treat each time instance as a measure, and the total of time instances/measures as the longitudinal k dimensional vector profile for each case. In turn, these trajectories can be combined (appended to one another) so that the cluster solution is based on similarities in evolution across all of the trace trajectories. For example, in our study of international health we appended the trajectory for per-capita GDP with the trajectory for longevity rates for each of the 156 countries in our database.

DOI 10.1002/cplx

TABLE 2

The Nine Clusters (Major and Minor Trends) for Allostatic Load

	Clusters[b]									
Factor/Components[a] Range (min − max)	**1**: Low Cholesterol	**2**: Healthy	**3**: High Blood Pressure	**4**: Low Stress Hormones	**5**: Metabolic Syndrome	**6**: High Blood Sugars	**7**: Low Stress Antagonist	**8**: High Stress Hormones	**9**: High Pro-Inflammatory Elements	ANOVA *F* test[c]
Stress hormones (−3.02 − 3.11)	−0.79[d]	0.33	0.35	−0.92	0.66	−0.22	−0.62	1.03	−0.30	118.41*
Metabolic syndrome (−2.81 − 2.90)	−0.55	−1.08	−0.40	0.16	1.22	1.00	−0.74	12	0.95	**177.97***
Proinflammatory (−3.03 − 3.08)	−0.41	−1.19	−0.71	0.29	0.99	0.57	−0.27	0.12	1.08	**154.72***
Cholesterol (−4.69 − 2.75)	−1.12	0.06	0.42	0.73	0.73	−0.01	−0.08	−0.69	−0.82	93.77
Blood sugars (−1.83 − 6.70)	−0.32	−0.48	−0.36	−0.13	0.18	3.71	−0.25	0.08	0.36	**215.42***
Stress antagonists (−3.86 − 2.26)	0.31	0.22	0.58	0.14	0.35	−0.10	−1.7	0.30	−0.73	102.06*
Blood pressure (−3.91 − 3.17)	−0.60	−1.10	0.94	−0.06	0.47	0.26	0.15	0.21	−0.52	80.78*
	N = 96[e]	*N* = 138	*N* = 155	*N* = 169	*N* = 144	*N* = 35	*N* = 109	*N* = 146	*N* = 104	

[a]These are the seven factors from Table 1, used to construct the different profiles for the nine clusters. Included below each factor is its min and max score possible, which comes from summing the biomarkers that loaded on it and converting this sum into a *z*-score.

[b]This 9-cluster solution was obtained using *k*-means, with standard Euclidian distance measures; convergence criterion was set to zero.

[c]Unstandardized F scores (ANOVA) demonstrating, for descriptive purposes only, the relative impact the seven factors had in determining cluster membership (* = *F* test was significant at .000. The factors with the three highest scores are highlighted).

[d]The score for each case, for each of the seven factors, was computed (as noted in "a" above) by summing each case's scores on the biomarkers for each factor, as shown in Table 1. In turn, these summed factor scores were converted into *z*-scores to normalize the data.

[e]Number of cases in each cluster.

2.2.3. Clustering to Corroborate

Third, it is unique in that we use k-means and the SOM as a method of corroboration. *k*-means is a partitional (as opposed to hierarchical) iterative clustering technique that seeks a single, simultaneous clustering solution for some proximity matrix. For *k*-means, reference vectors are centroids, representing the average for all the cases in a cluster. The SOM is a topographical artificial neural network that maps high-dimensional data onto a smaller, three-dimensional space, while preserving, as much as possible, the complex patterns of relationships amongst these data. For the SOM, reference vectors are actual points, neurons, which represent the weighted average of the cases clustering around it. Both *k*-means and the SOM are forms of unsupervised learning, as cluster membership is not known ahead of time.

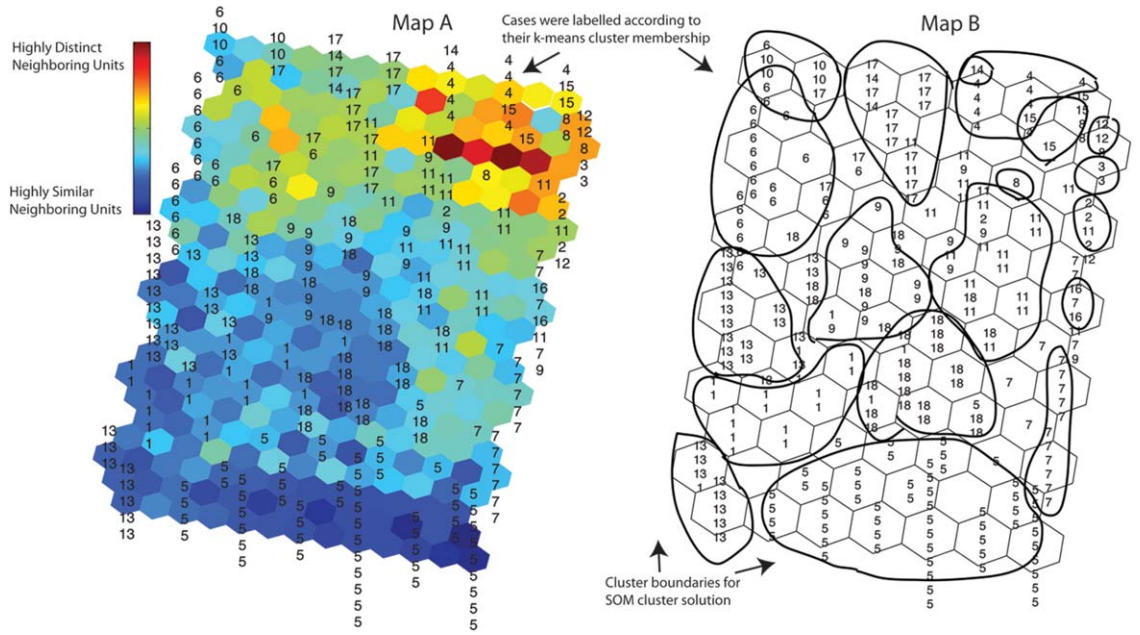
In terms of a case-based density approach, these methods are used in combination as follows: *k*-means is used first because it requires that the number of centroids be identified ahead of time, based largely on some rationale, even if tentative or conjectural. As shown in Table 2, following convention (and as discussed earlier), the goal of multiple runs is to find a solution that fits the data well and resonates with our theoretical model, even when exploratory.

Next, the SOM is run. Because the SOM is entirely unsupervised, if it arrives at a solution similar to the k-means this provides an effective method of corroboration. The closer the final quantization error and final topographic error are to zero, the better the fit of the model.

The SOM graphs its cluster solution onto a variety of three-dimensional, topographical maps. The three we typically use are the *u*-matrix, eigenvector map, and components map. On the *u*-matrix and eigenvector maps, cases most like one another are graphically positioned as nearby neighbors, with the most unlike cases placed furthest apart. Both maps are also topographical: valleys, or darker colored, areas are more similar in profile; while hilly, or brighter colored areas, are more distinct. The component maps (which we will discuss in Step 10) visualize how each of the variables (traces) from the complex profile of study contribute to the final cluster solution and to the positioning of cases on the *u*-matrix and eigenvector map.

DOI 10.1002/cplx

FIGURE 6



The eighteen cluster solution for depression/physical wellbeing study.

A good example of this output comes from our study on depression and physical wellbeing. As Shown in Figure 6, Map A and Map B are graphic representations of the cluster solution arrived at by the Self-Organizing Map (SOM) Neural Net, referred to as the *U*-Matrix. In Figure 6, Map A is the three-dimensional (topographical) *u*-matrix: for it, the SOM adds hexagons to allow for visual inspection of the degree of similarity amongst neighboring map units; the dark blue areas indicate neighborhoods of cases that are highly similar; in turn, bright yellow and red areas, as in the upper right corner of the map, indicate cases that are very different from the rest. Map B is a two-dimensional version of Map A that allows for visual inspection of how the SOM clustered the individual cases. Cases on this version of the *u*-matrix (as well as Map A) were labeled according to their *k*-means cluster membership (the 18 cluster solution we discussed earlier) to see if the SOM arrived at a similar solution, which (roughly speaking) it did.

2.3. Clarification of the Difference between Steps 2 – 5 and Steps 6 – 8

Before proceeding, and as a point of clarification, in steps 2–5 we are still dealing with data and there are no dynamic models involved yet. In steps 6–8 we are building the microscopic vector field f to use in the ODE model [Eq. (1)] to capture the microscopic movements of individual trajectories and in the PDE model [Eq. (2)] to capture the macroscopic trends in the form of motion of densities. In other words, it is in Steps 6–8 that we are doing the actual functional modeling using genetic algorithms and curve fitting algorithms, which we explain below. Hence, steps 6–8 deal with the actual model building process i.e the microscopic and macroscopic models in the form of the vector field ODE and the density PDE.

2.4. Microscopic and Macroscopic Models

To construct our microscopic model (Steps 6 and 7), we employ a combination of genetic algorithms and ODEs; and to construct our macroscopic model (Steps 8 and 9), we employ the continuity (advection) PDE in application to the vector field generated by our microscopic model. As such, before moving on to our next set of steps, a bit of detail on the mathematics behind them is necessary – for a complete explanation, see [4]. Before proceeding, however, just to make clear: we use the longitudinal trajectories of the traces; to which we fit a microscopic vector field f; from here we used the advection PDE (see below) to simulate the motion of a density of initial conditions $\rho_0(x)$ i.e. the macroscopic movement $\rho(x, t)$ of an ensemble of initial condition of our choice—which is usually motivated by the problem at hand as well as the data—to investigate the macroscopic movement.

Here, then, are the microscopic and macroscopic models upon which our approach is based:

DOI 10.1002/cplx

Microscopic model for nonlinear evolution of each case trajectory

$$\Downarrow$$

$$x' = f(x); x(0) = x_0; x \in X \subset \mathbb{R}^K \tag{1}$$

Macroscopic model for the linear evolution of densities of cases

$$\Downarrow$$

$$\rho_t + \nabla \cdot (\rho f) = 0; \rho|_{\Gamma_i} = 0; \rho(x, 0) = \rho_0(x). \tag{2}$$

2.5. Steps 6 and 7: The Microscopic Model and Steady-State Behaviors

In addition to the mathematics upon which they are based, Step 6 and Step 7 involve a rather complicated set of procedures, which we have outlined in detail elsewhere. Our goal here is to provide a quick overview of the procedures involved in completing them. For more details see [4,5].

2.5.1. Identify the Data-Driven Vector Field

In terms of constructing the microscopic model, the form of the vector field f, which is a part of the ODE (1), is completely unknown. In other words, we do not have a preconceived function for the vector field model. As such – and for a second time – we employ our knowledge-free approach to modeling: this time looking for the best fit among polynomials of arbitrary degree using genetic algorithms.

To run our genetic algorithm, we used Eureqa's software http://formulize.nutonian.com. The component functions of the vector field are constrained to have a polynomial form i.e. powers of trace variables with addition, subtraction, multiplication, and constants. We generally choose a polynomial fit without any constraint on the degree, and use the mean squared error with the Akaike information criterion as a measure of error. The software provides a measure of stability and maturity, where 'stability' refers proportionally to how long ago the top solutions were modified among the multiple solutions provided; and where 'maturity' refers to how long ago any of the solutions have improved. Stability and maturity values close to 100 percent mean that the solutions cannot be improved any more. The software shows multiple solutions ordered according to their level of complexity of polynomials and level of fit. The top solutions are extremely complicated polynomials with a very good level of fit (i.e., lowest error), whereas the bottom solutions are extremely simple and thereby giving the worst error. The mid-range solutions are the best in terms of complexity of polynomial terms and error fit.

2.5.2. Validity Check

When stability and maturity are close to 100 percent, this indicates that not much improvement in the error is going to happen. The top to mid-range solutions are copied as seeds and the algorithm is rerun to obtain a refinement of solutions. The error values are examined to ensure they are to the order of $1e-4$ or lesser using the stability and maturity approach. For more information please go to http://formulize.nutonian.com/documentation/eureqa/.

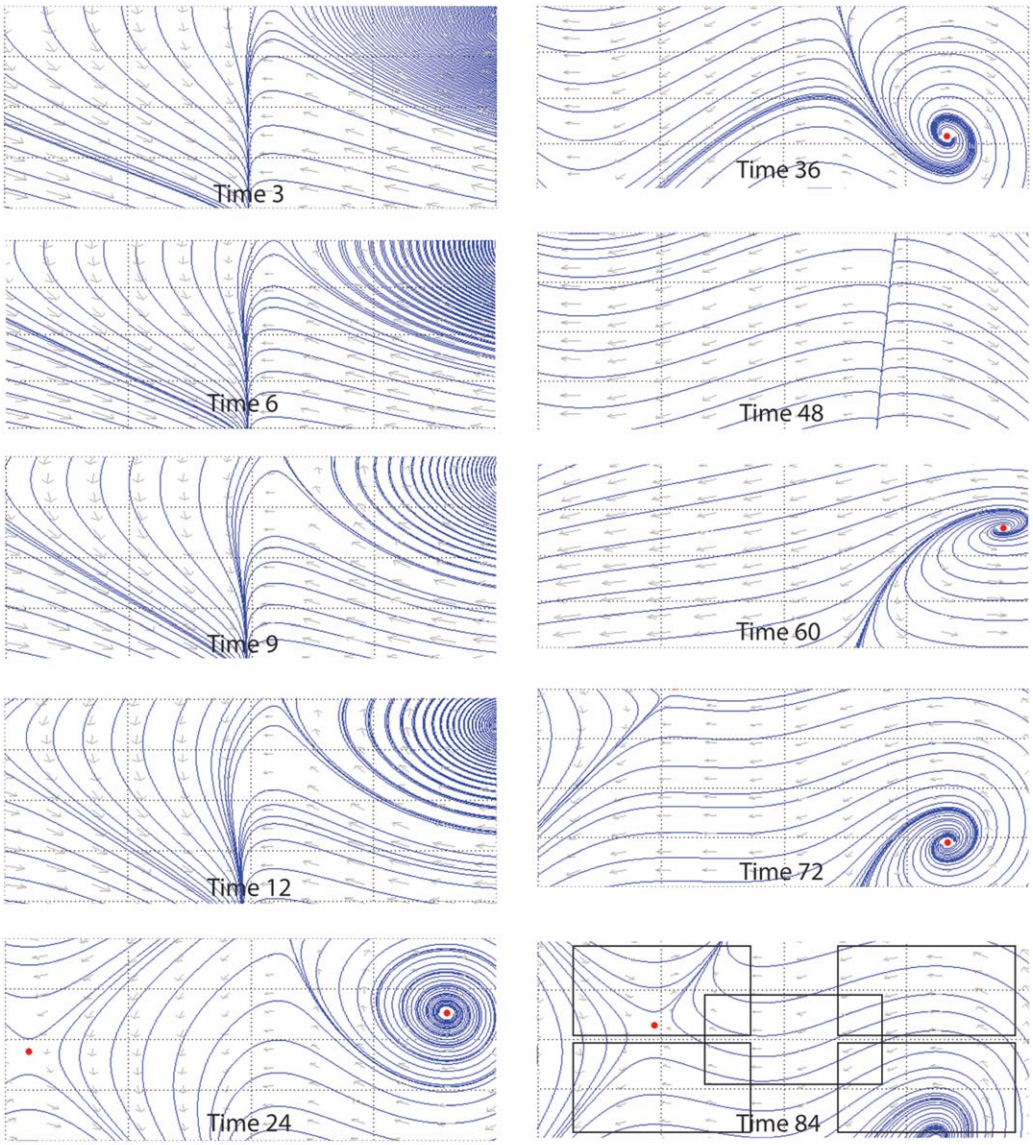
The genetic algorithm also allows for other error criteria and other kinds of 'function fitting' ranging from trigonometric to hyperbolic, rational, exponential and other complicated functions, but we chose polynomials because they are known to be dense in many complicated function spaces and are easier to handle when used as component of a vector field for an ODE. They are also known to capture well most complex phenomena such as chaotic attractors, etc. Experimentation with different error criteria often shows the minimum mean squared error (with the Akaike Information Criterion) between the model and the velocity data gave the lowest error values of less than $1e-4$. For a list of error metrics available please visit http://formulize.nutonian.com/documentation/eureqa/general-reference/error-metrics/and for a list of functions available to fit, please visit http://formulize.nutonian.com/documentation/eureqa/general-reference/building-blocks/. Our attempt to fit a curve to this data-driven vector field constitutes another of the novel aspects of our approach. We chose polynomials since (a) they are known to be dense in a variety of complicated functional spaces, and (b) they are easier to simulate when used as vector fields in ODEs.

2.5.3. Obtaining the Microscopic Model

To obtain a vector-field model for the velocities of the traces, we use a curve fitting algorithm to fit each case trajectory with a smooth curve; which we then differentiate with respect to time. MATLAB software was used to fit piecewise cubic Hermite interpolant polynomials (to minimize overshoot and oscillation) using the *pchip* command. The *phcip* interpolant is also known to be" shape preserving" and known to respect "monotonicity" in addition to being less expensive to set up numerically, and hence the reason for our choice. The first derivative is also known to be continuous and hence it is easier to differentiate the interpolant and evaluate the derivative of the function using the *fnder* command in MATLAB. Pictures of the actual trajectory and the Hermite interpolant were also used to visually make sure that the fit was good. Since the fit is an interpolant, the value of the trajectory at the known instants of time are exactly matched along with a smooth first derivative due to the nature of the Hermite basis functions. For an example demonstration please go to http://in.mathworks.com/help/matlab/ref/pchip.html. Our purpose in doing so is to find the instantaneous (continuous) velocities for the time instants provided by the data—which we discussed above in Step 2. A discrete

DOI 10.1002/cplx

FIGURE 7



Microscopic model for depression and physical wellbeing study.

velocity vector $f(x_k)$ is thus obtained, situated at each of the cases x_k.

2.5.4. Searching for Complex Steady-State Behaviors

The utility of the microscopic model is that, unlike the cluster model, it is devoid of cases, constituting, instead, the data-driven space of all possible trajectories. Equally important, it can be visualized as a movie across all instantaneous time-stamps in the database. As such, the model can be visually inspected to identify important steady-state behaviors and to note the manner in which the trajectories evolve across continuous time, including changes in velocity.

For example, Figure 7 shows the state-space for our depression and physical wellbeing study, which included 84 monthly time-stamps across a 7-year period of time. Ten time-stamps are shown, beginning with four time-stamps from the first year (3 months, 6, months, 9 months, and 12 months) and then one time-stamp for

DOI 10.1002/cplx

each subsequent year, each constituting the point at which new data were collected. Looking at the time-stamps, one sees (1) the emergence of stable spiraling equilibrium points marked in red, which enter the relevant regions of the state space for time-stamps 3 through 24; (2) an unstable spiraling equilibrium point appearing and leaving the state space between time-stamp 36 and 84; and (3) a saddle appearing at time-stamp 84.

Because, in our approach, the ODE that models the evolution of depression trajectories is non-autonomous $\dot{x}=f(x,t)$, time t is an independent variable and hence the vector field changes it nature as time evolves, as seen in the emergence and disappearance of various steady-state and transient behavior (such as rifts etc.). This is one of the main advantages of using an ODE to model the trajectories: both the steady-state and transient behaviors of trajectories can be studied with time as an independent variable (which allows for change of these behaviors across time) by using the multiple ODE models (both autonomous and non-autonomous) that the genetic algorithm fits as possible explanations of the trajectories from the standpoint of ODEs.

2.6. Steps 8 and 9: The Macroscopic Density Model

With the cluster and microscopic models complete, the next stage is to assemble the macroscopic model, which involves two key steps.

2.6.1. Simulating the Transport of Densities

First, we take the vector field f from our microscopic model—which is governed by the ODE—and use the advection PDE to translate it into the macroscopic motion of case-based densities. In doing so, we add a third level to our approach, focused on the macroscopic, nonequilibrium properties of the system as a whole.

Our approach is motivated by thermodynamics [25], where the state of the system is the characteristic of a density of particles and their properties (as in the case of temperature or pressure) rather than the individual particles themselves. As we discussed earlier, the major challenge of modeling longitudinal data is that trajectories are often highly complex. For example, in our study of international health, our microscopic model – while useful for identifying major and minor trends and steady-state behaviors – was still incredibly dynamic (Figures 3 and 7). In such instances, our approach is useful because, Haken [26], it models the ensemble of cases as the macroscopic movement of densities across continuous time; which are, generally speaking, lower dynamic and therefore easier to model for common patterns and trends across time.

The key aspect of the advection PDE is that it models the transport of a physical quantity according to a given vector field f (as in the case of our microscopic model) in addition to conserving the physical quantity itself. The dynamical state of the advection PDE is a density ρ, which is a function of both space x and time t. In turn, the density function ρ is basically the physical quantity per unit area in two dimensions.

Given an initial distribution of cases $\rho_0(x)$, the advection PDE simulates the evolution of ρ_0 under the assumption that (a) the total number of cases remains the same, and (b) each case x moves according to the velocity vector $f(x,t)$. The boundary condition $\rho|_{\Gamma_i}=0$ ensures that no new cases enter the state space through the inflow portion of the boundary given by

$$\Gamma_i=\{x \in \partial X : f(x)\cdot \eta < 0\}, \tag{3}$$

where η stands for the outward normal on that boundary ∂X. And, we use the vector field f obtained in microscopic model to simulate the advection equation given in (2) along with the boundary condition (3). As a final note, the validity check for the motion of densities is mathematically inherent because the vector field f is already checked in the microscopic model. (For more details on our approach, see [4,5].)

2.6.2. Simulating Initial and Novel Conditions

With the macroscopic model built, the next step (Step 9) is to run it—which we do by introducing known or novel sets of initial conditions. Unlike the microscopic model, which is devoid of cases, the macroscopic movie reintroduces different distributions of cases back into the model to explore their actual trajectory amongst all possible trajectories in the microscopic model.

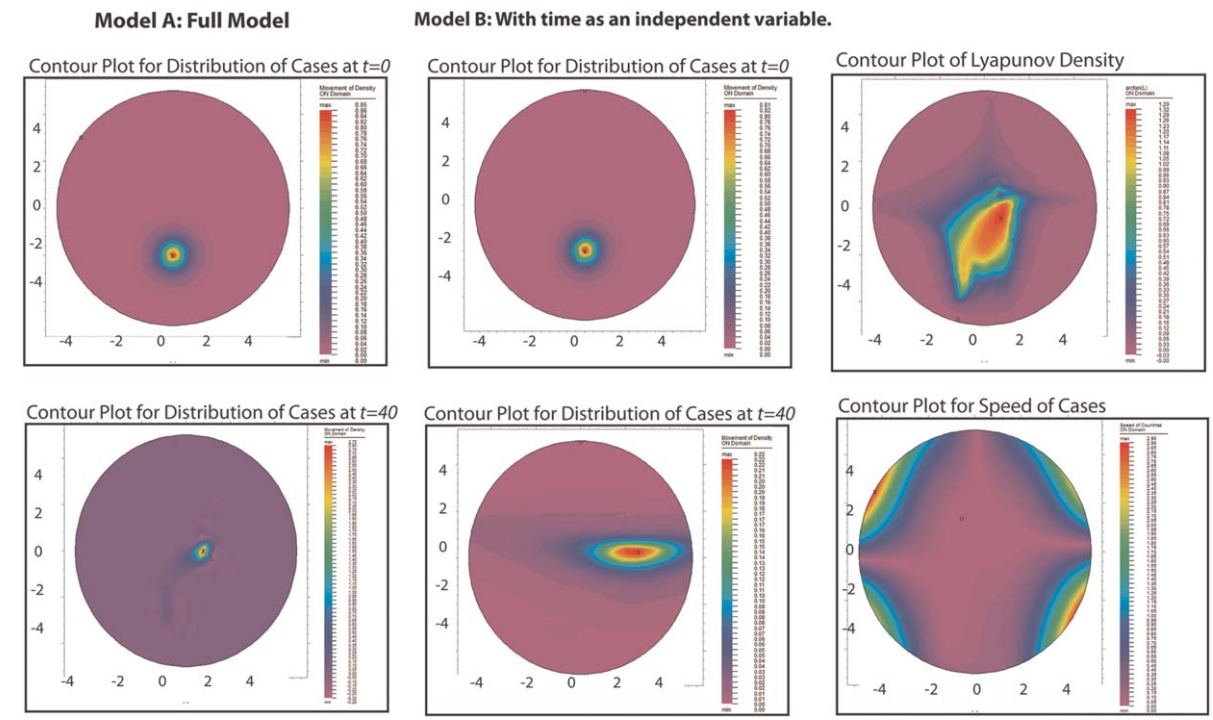
For us, these initial conditions come in three types: (1) conditions based on the initial dataset upon which the microscopic vector field f was based; (2) conditions based on the major and minor trends identified in the cluster and microscopic models; and (3) novel conditions researchers wish to explore, based on the results from running conditions (1) and (2).

Simulating these 'various' initial conditions is important to our approach because it brings us full circle, moving us from the inverse to the forward problem in physics: in other words, while we use the microscopic model f to simulate the macroscopic evolution of densities (using the advection equation), we do so by returning to the initial conditions of the raw data (be it known or novel) for corroboration.

For example, Figure 8 provides a series of snapshots from a simulation we made for our international health study. In terms of reading Figure 8, the x-axis represents GDP and the y-axis represents life expectancy; also, scores on the axes were converted to z-scores for normalization and comparison. In this example, the initial conditions $t=0$ (which are shown in Models A and B) were based on the original Gapminder dataset http://www.gapminder.

DOI 10.1002/cplx

FIGURE 8



Macroscopic model of international health study.

org/. (The complete movie can be found at http://www.personal.kent.edu/~bcastel3/macroscopic_model.mp4.)

In contrast, Figure 9 provides a series of snapshots from a simulation we made for our depression and physical well-being study. In this simulation, five different sets of initial conditions we explored, based on key trends identified in the cluster and microscopic models. As with Figure 8, scores were converted to *z*-scores, with the *y*-axis representing physical wellbeing; and the *x*-axis representing depression.

As these two examples illustrate, our macroscopic model adds several advantages to our multi-level, case-based approach to modeling complex health trajectories.

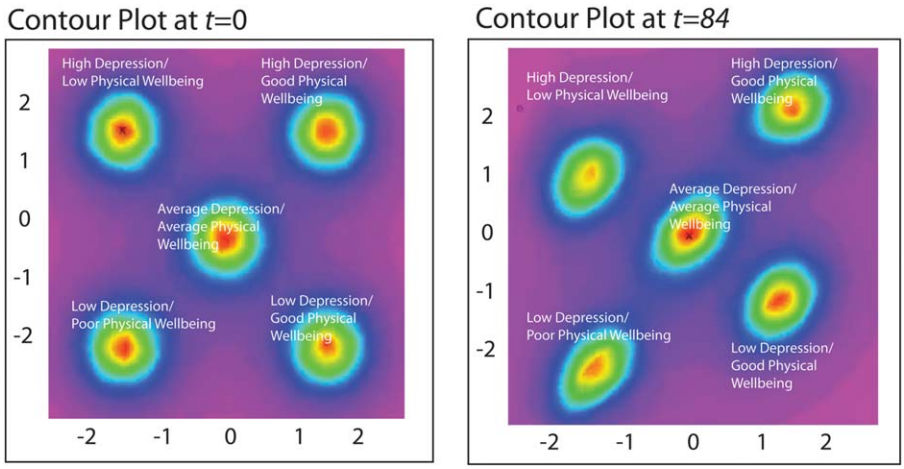
1. To begin, different regions of the simulation can be explored to see how different sets of cases evolve (speed up, slow down, spread out, condense inward toward the center of the density, etc) across time.
2. And, these movements can be calibrated using a number of indicators, such as the contour plot of speed (magnitude of velocity)—as shown in the lower right graph in Figure 8.
3. Also, the nonequilibrium clustering of trajectories during transient times can be studied by looking at the Lyapunov density plot. In Figure 9, for example, high values in the Lyapunov density plot (shown in the upper right graph) indicate that a large number of trajectories have squeezed through that region in the state space.
4. We can also use these simulations to predict the longitudinal evolution of cases across time and space.
5. And, we can study the complexity of various transient case dynamics, which we do by stopping the evolution of the model prior to some key moment in the simulation.
6. Also, based on the exploration of various novel conditions, predictions can be made for the evolution of case profiles and time instants that are not part of the database.
7. And, multiple models can be tested simultaneously to find the model that best explains the data.
8. Finally, new data can be incorporated with ease into the modeling process, thereby providing us with a means to improve the model's fit and predictive value in response to the database's evolution, expansion, development, etc.

2.7. Step 10: Constructing Multiple Accounts

Consistent with the overarching theme of case-based complexity—which seeks to find differences through idiographic, case-comparative analysis—our approach, once again, distinguishes itself from convention. In Step 10, we

DOI 10.1002/cplx

FIGURE 9



Macroscopic model for depression and physical wellbeing study.

do not seek to build a single causal model or overarching account of the data. Instead, we seek to construct multiple models, multiple accounts. And, we aim these multiple accounts at explaining (exploring, understanding, etc.,) key differences, (distinctions, variations, nuances etc.,) in the data.

Equally important, we assume that these multiple accounts come from the theoretical map—as constructed in Step 4—and its k dimensional profile of traces. And we assume—as discussed in Step 1—that this map and profiles are best studied as complex systems. As such, while it is useful to explore variable-based trends across data; in our approach the emphasis is on the intersection of traces and their self-organizing and emergent impact on some initial (dependent/outcome) trace of concern.

With our map and case-based profile in hand, we construct our multiple accounts by employing the following two-stage process:

2.7.1 Corroboration of the Three Models

During the course of completing steps one through nine, a significant amount of information is generated. It is therefore necessary, as a first course of action, to summarize and further corroborate these data into a *multilevel, working narrative* to identify key issues for which we seek to develop an account.

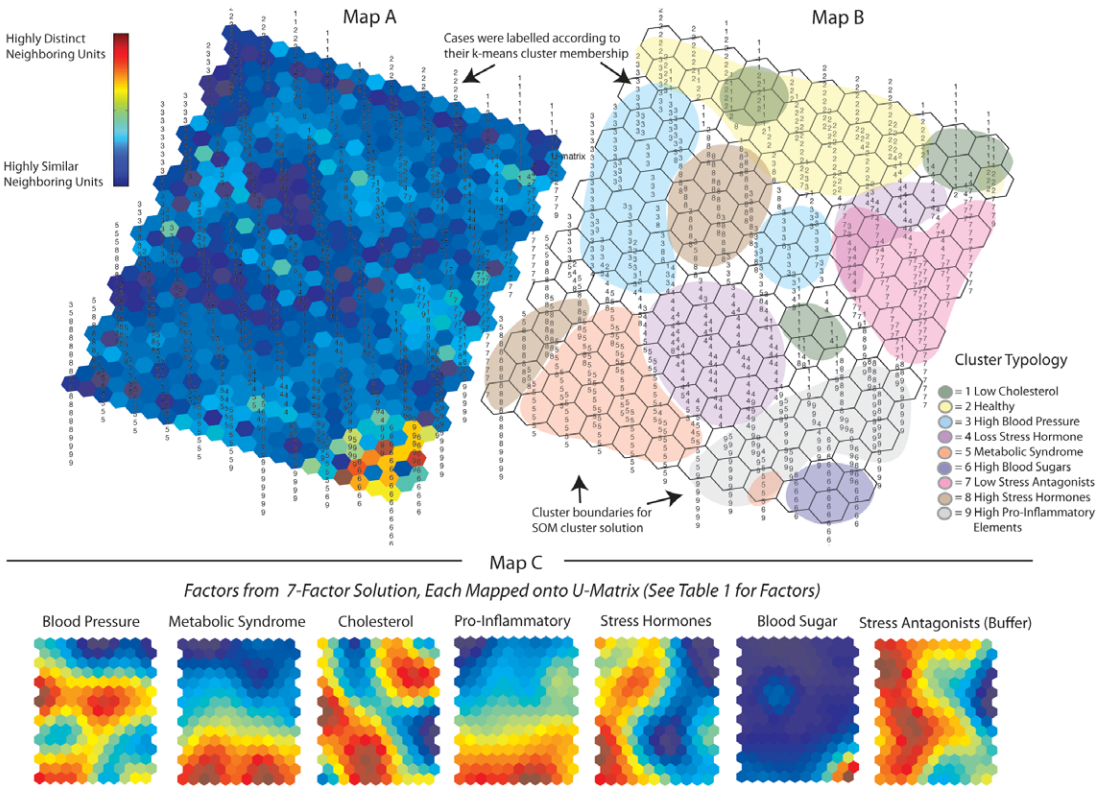
1. **The Cluster Model**: An easy place to start is with the cluster model. Here, the goal is to verify further the veracity of the cluster trajectories and the major and minor trends they represent. Such verification includes working with context experts to:
    1. determine if the clusters make empirical or theoretical sense.
    2. examine if certain clusters need to be discarded or combined to create a larger cluster.
    3. determine – as discussed in Step 2.2.1 – if further subclustering is necessary or empirically or theoretically meaningful.
    4. name the final cluster and subcluster solutions, as well as major and minor trends, based on how their trajectories differ from one another.
    5. assemble all this information into a working narrative.
    6. And, finally, identify key issues for which an account of this narrative is required, including hypothesizing how the other k dimensional traces in the theoretical model and case-based profile might account for these differences.

For example, in our study on depression and physical wellbeing, the two lead clinicians (both authors on the current paper) had to pore over the data to make sure everything made empirical sense, including tentatively hypothesizing how the other traces in the theoretical model might account for these differences.

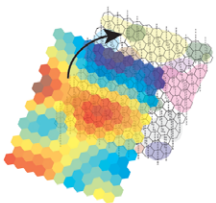
The same was true of our allostatic load study, albeit at an even greater level of detail, as the clusters for this study were based on our factor analysis (we discussed this earlier); which was, in turn, based on our 20 key biomarkers. As such, the biologists and clinicians on our team had to corroborate the biomarker linkages found in our cluster solutions, shown in Table 2, to make sure they fit with what they and the literature knew to be biologically true or possible.

DOI 10.1002/cplx

FIGURE 10



Map C shows how each factor is distributed across the u-matrix -- the more red the higher the value; the more blue the lower the value



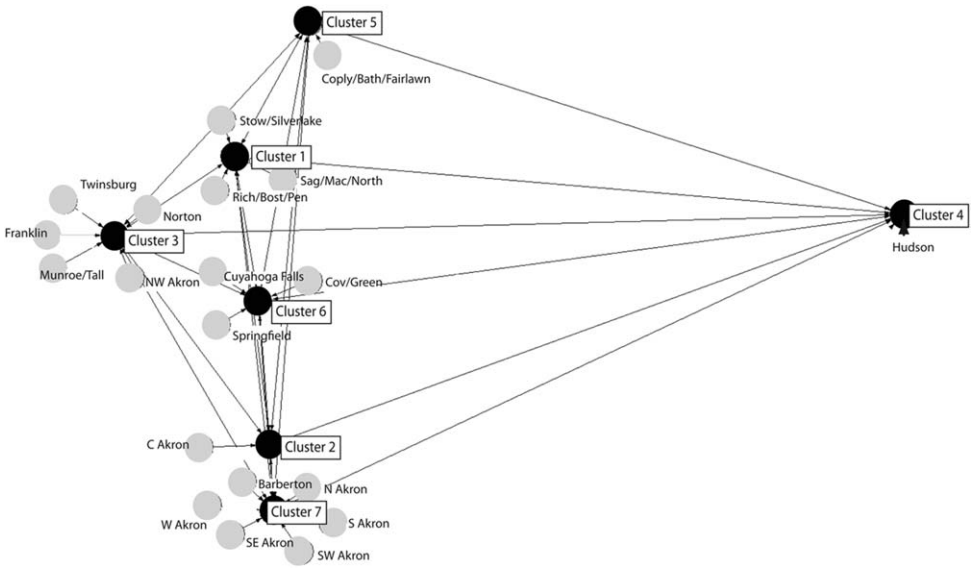
U-Matrix and Components Maps for Nine Allostatic Load Profiles: *Map A* and *Map B* are graphic representations of the cluster solution arrived at by the Self-Organizing Map (SOM) Neural Net, referred to as the *U-Matrix*. In terms of the information they provide, *Map A* is a three-dimensional (topographical) u-matrix: for it, the SOM adds hexagons to the original 15X11 map to allow for visual inspection of the degree of similarity amongst neighboring map units; the dark blue areas indicate neighborhoods of cases that are highly similar; in turn, bright yellow and red areas, as in the lower right corner of the map, indicate highly defined cluster boundaries. *Map B* is a two-dimensional version of *Map A* that allows for visual inspection of how the SOM clustered the individual cases. Cases on this version of the u-matrix (as well as *Map A*) were labelled according to their k-means cluster membership (The 9 cluster solution showin Table 2) to see if the SOM would arrive at a similar solution. *Map C* is a graphic representation of the relative influence that the seven factors (shown in Table I) had on the SOM cluster solution. The SOM generates a mini-map for the seven factors, each of which can be overlaid across maps A and B. Each of these mini-maps can then be inspected visually to examine what its rates are across the different neighborhoods (clusters of cases). Dark blue areas indicate the lowest rates for a factor; and the bright red areas indicate the highest rates for a factor. For example, looking at the mini-map for Factor 6 (*Blood Sugar*), its rates are extremely low across most of the map, except for the lower right corner, which is where (looking at *Map A* and *Map B*) the SOM placed Cluster 6.

The nine clusters for allostatic load as mapped by the SOM.

In turn, the nine clusters shown in Figure 10 also had to make sense in terms of our theoretical model of allostatic load (See, from earlier, Figure 5). For example, Map C in Figure 10 is a graphic representation of the relative influence that the seven traces (shown in Table 2) had on the SOM cluster solution. The SOM generated a mini-map for the seven traces, each of which can be overlaid across maps A and B. Each of these mini-maps was also visually inspected to examine what its rates were across the different neighborhoods (clusters of cases, See Map B, Figure 10). Dark blue

DOI 10.1002/cplx

FIGURE 11



The seven cluster network map for the Summit county community health study.

areas in Map C indicated the lowest rates for a factor; and the bright red areas indicated the highest rates for a factor. For example, looking at the map for Factor 6 (blood sugar), its rates were extremely low across most of the map, except for the lower right corner, where (looking at Map B) the SOM placed Cluster 6.

2. **The Microscopic Model**: Next, the insights from the cluster model need to be integrated and further corroborated with the microscopic model. Such integration and corroboration includes working with context experts to:
    1. verify the empirical and theoretical utility of the various temporary equilibrium behaviors (temporary because the behavior exists for one time instant and changes for later instants because of the non-autonomous nature of the vector field) initially identified during Step 2.4.4 of the model building process.
    2. use the cluster trajectories to visually corroborate the manner in which the microscopic vector field evolves across continuous time.
    3. examine how key velocity change moments in the microscopic model coincide with key changes in the trajectories of various major and minor clusters. This is done by (a) identifying the region in the state-space where the behavior occurs; (b) studying the clusters corresponding to the regions; and (c) watching the movies and the actual trajectories from data (both overall and at the interesting time instants) to see if they corroborate one another.
    4. use the findings from the cluster and microscopic model to identify different regions of the state-space that are qualitatively important, and to then label them accordingly.
    5. and, finally, hypothesize how the other k dimensional traces in the theoretical model and case-based profile might account for these differences in the state-space and their respective evolution across time-space.

For example, as shown in Figure 7, in terms of our study of depression and physical wellbeing, we identified five major regions into which we could fit all 18 of our clusters, based on the major and minor trends we noted in our cluster solution. Roughly speaking, these regions approximated how physical wellbeing and depression work together as a way for sorting clusters according to their key differences. Figure 9 – which we already mentioned but will discuss further in a moment – provides the name of these five major regions. In addition, we confirmed that the temporary equilibrium behaviors from our microscopic model (Figure 7) coincided with key shifts in the cluster trajectories from our cluster model.

3. **The Macroscopic Model**: The last part of our corroboration process is to add the macroscopic model to our multilevel narrative by integrating the insights from its density simulations. Such

DOI 10.1002/cplx

integration and corroboration includes working with context experts to:

1. confirm (both empirically and theoretically) the utility of the results gained in Step 2.5.2 of the model building process, which includes validating the various novel scenarios simulated, based on the results from the microscopic and cluster models.
2. verify how and also when the temporary equilibrium behaviors identified in the microscopic model manifest themselves in the density model, given that both models are based on the same vector field generated by the ODE.
3. and, finally, hypothesize how the other k dimensional traces in the theoretical model and case-based profile might account for these differences in the density model.

For example, looking at Figure 9 from our depression and physical wellbeing study, we see two snapshots, $t = 0$ months and $t = 84$ months, which we labeled according to the results from our cluster and microscopic models. During the course of exploring these five simulations, one of the findings (amongst many) that stood out was how the density plot for the healthiest region (lower right) changed across time, moving more toward the center. We also noted how this 'healthier' plot shifted and became more distributed across time, with the upper right side stretched more toward increased depression, and the lower right stretched more toward decreased physical wellbeing. The question we sought to answer, based on this finding, and for which we engaged in a series of hypotheses, was why? Answering this question takes us to the final stage of our model building process: constructing multiple accounts.

2.7.2. Constructing Multiple Accounts

The final stage of the modeling process is to construct a series of accounts (causal models) that help make sense of the complex health trajectories studied. Again, no one account is expected to explain everything. Instead, we seek multiple accounts, multiple explanations.

To do so, we begin with method. Unlike the previous steps, however, this last stage need not follow any particular methodological protocol. In fact, in our work we have employed a variety of methods, including statistical, qualitative and historical analysis. Of these methods, however, perhaps the most useful is to rerun steps five through nine.

As a quick overview: starting with the cluster model, one would proceed, as required, to examine a new set of traces. However, this time there is an added dimension, as the goal is to construct an account of how the nonlinear dynamics of these additional traces coincide, interconnect with, influence or impact the original outcome traces. For example, in our depression and health study, we were specifically interested in how the complex intersection of employment, income and negative life events impact—across time—our 18 cluster trajectories, slowly stitching these new traces together to form a complex model of these data.

Nonetheless, often one does not have the luxury of continuous data, or researchers may be concerned with traces of a different type, as in the case of discrete, qualitative or historical data, or they may be interested in other methods, as in the case of network analysis.

For example, in our public health study, we used multiple linear regression and the unstandardized F-scores from our k-means cluster analysis to determine the relative impact that various compositional and contextual traces had on our community-level health outcomes.

In this same study we also used a combination of qualitative and historical data. For example, to examine the views and opinions of people living in the poorer communities surrounding the two urban centers in our study, we turned to a series of focus groups that local public health researchers had done. And, to make historical sense of how out-migration impacted community-level health, we turned to a local newspaper series on access to healthcare.

Going even further, we employed the tools of complex network analysis, examining (pace Christakis and Fowler's work on obesity networks [27]) how changes in communities in one part of our county-wide network influenced changes in another. Figure 11, for example, is a network representation of our cluster analysis data. The network is made up of the seven clusters we identified in our study, labeled 1 through 7. Around each cluster are the communities associated with it.

In terms of reading this network, the greater the distance between cluster centers, the less alike these clusters are in health and economic wellbeing; and, the greater the distance a community is from its cluster center, the less similar its configuration is to the other communities in its cluster. One of the questions we examined using this network was: given changes in the overall economic wellbeing of all twenty communities, which of the poorer communities fell further into a poverty trap, relative to the rest? The answer was found in the bottom two clusters, both 2 and 7, with the communities in Cluster 7 having the poorest, overall, health outcomes.

Still, despite these differences in method and technique, when it comes to the final step of our case-based density approach, two things are consistent in the account building process. The first has to do with focus: no matter what the technique used, the purpose is to construct multiple accounts that help explain the nonlinear dynamics of complex longitudinal health data. More specifically, this means making sense of:

DOI 10.1002/cplx

1. the different cluster trajectories;
2. major and minor trends;
3. various transient, steady-state behaviors;
4. the movement (across space/time) of various density distributions;
5. and, the evolution and dynamics of different regions of the state-space, including:
    a. differences in the speed and velocity of key trajectories; and
    b. the prediction of different known or novel density outcomes.

The second has to do with the preliminary division of the data. Regardless of the method used, before one constructs any account, the data need to be divided according to each case's cluster membership and the corresponding trend or region of the state-space to which it belongs. In other words, it is necessary to divide the database into separate sub-databases according to the different clusters or trends to which cases belong. Once divided, one can then explore these clusters and trends separately and in comparison to one another, looking for different patterns within and also across clusters and trends.

With the construction of these different and multiple accounts complete, one has reached the end of the modeling process, that is, unless, as a function of the study, new or novel additional traces or cases are included, and therefore further modeling is required.

3. CONCLUSION

In this paper, we have brought together the various steps outlined in our previous research [4,5] to organize them into a single method, called a case-based density approach (CBD),which involves ten steps.

In terms of summarizing these ten steps, our main thesis is that complex longitudinal data are inherently multilevel, case-based systems that manifest themselves from the bottom-up (as the microscopic, high-dynamic behavior of individual cases) as well as the top-down (as the macroscopic, low-dynamic behavior of densities).

Our secondary thesis is that to model such complex case-based, longitudinal data, researchers need to acknowledge that singular one-size-fits-all models are not sufficient; instead, new and multiple models are necessary. Furthermore, these multiple accounts need to be data-driven and predictive (if only in the short range).

Third, we acknowledge that complex phenomena cannot be perfectly (or often even directly) modeled using mathematical models. Instead, one typically studies the traces of a system's complexity. Related, these traces are often unknown; and only identified through multiple exchanges with subject matter experts. However, modeling the traces of complexity is the first step towards understanding a system's causal mechanisms, which is what we have endeavored to achieve with our approach.

Given these three main points, to model complex health trajectories it is necessary to draw upon a wide variety of concepts and techniques from across the complexity sciences, including the ideas of nonequilibrium statistical mechanics, transport theory and thermodynamics. To model the motion of density of cases, we specifically employed the advection PDE, which serves as a conduit to translate the microscopic motions modeled by the vector field f into the macroscopic evolution of the density ρ, while preserving the number of cases intact. This is a very novel feature of our approach, which has yet to be used to observe the complex behavior of health trajectories longitudinally in time.

Finally, Step 10 is a very important part of the modeling process, as it seeks to provide a series of accounts of the causal mechanisms that drive the nonlinear dynamics of health trajectories. It is crucial during this stage that the results do not just show the experts what they want to see. In other words, the 'knowledge-free' part of our approach is crucial, as it also helps to challenge the preconceptions of the content experts on the study. As such, and as mentioned earlier, we are currently studying a large sample of the Diamond Prospective Longitudinal Cohort Study to develop further this step: our goal is to push the analysis into new and novel insights that challenge current views on the topic. This is the future work that is currently underway.

REFERENCES

1. Castellani, B.; Rajaram, R.; Buckwalter, J.G.; Ball, M.; Hafferty, F.W. Place and Health as Complex Systems. Springer Briefs on Public Health, Germany, 2014.
2. Feng, Q.Y.; Griffiths, F.; Parsons, N.; Gunn, J. An exploratory statistical approach to depression pattern identification. Physica A: Stat Mech Appl 2013, 392, 889–901.
3. Gunn, J.; Elliott, P.; Densley, K.; Middleton, A.; Ambresin, G.; Dowrick, C.; Herrman, H.; Hegarty, K.; Gilchrist, G.; Griffiths, F. A trajectory-based approach to understand the factors associated with persistent depressive symptoms in primary care. J Affect Disorders 2013, 148, 338–346.
4. Rajaram, R.; Castellani, B. Modeling complex systems macroscopically: Case/agent-based modeling, synergetics and the continuity equation. Complexity 2012, 18, 8–17.
5. Rajaram, R.; Castellani, B.: The utility of non-equilibrium statistical mechanics, specifically transport theory, for modeling cohort data. Complexity 2014, 20, 45–57.
6. Castellani, B., Hafferty, F.: Sociology and Complexity Science: A New Field of Inquiry. Springer, Germany, 2009.

DOI 10.1002/cplx

7. Castellani, B.; Rajaram, R. Case-based modeling and the sacs toolkit: A mathematical outline. Comput Math Organ Theory 2012, 18, 153–174.
8. Byrne, D.; Callaghan, G. Complexity Theory and the Social Sciences: The State of the Art. Routledge: UK, 2013.
9. Byrne, D., (eds.), C.C.R. The Sage handbook of case-based methods. Sage Publications Ltd: London, UK, 2013.
10. Buckwalter, J.; Castellani, B.; McEwen, B.; Karlamangla, A.S.; Rizzo, A.A.; John, B.; O'Donnell, K.; Seeman, T. Allostatic load as a complex clinical construct: A case-based computational modeling approach. Complexity 2015, (in review).
11. Rajaram, R.; Castellani, B.; Gunn, J.; Uprichard, E.; Byrne, D.; Griffiths, F. Modeling depression and physical wellbeing longitudinally: A case-based density approach. Am J Public Health, submitted.
12. Byrne, D. Complex realist and configurational approaches to cases: A radical synthesis. Case-Based Methods, 2009, 101–111.
13. Ragin, C.C. Casing and the process of social inquiry1. What is a case?: Exploring the foundations of social inquiry. 1992, 217.
14. Ragin, C.C., Becker, H.S. What is a Case?: Exploring the Foundations of Social Inquiry. Cambridge University Press: Cambridge, UK, 1992.
15. Cilliers, P. Boundaries, hierarchies and networks in complex systems. Int J Innovat Manag 2001, 5, 135–147.
16. Mitchell, M. Complexity: A Guided Tour. Oxford University Press: New York, USA, 2009.
17. Sornette, D.: Probability distributions in complex systems. In: Encyclopedia of Complexity and Systems Science, R. Meyers, Ed.; Springer: New York, USA, 2009; pp 7009–7024.
18. Kuo, B.C., Golnaraghi, F. Automatic Control Systems. Wiley: USA, 2002.
19. Jain, A. Data clustering: 50 years beyond k-means. Pattern Recogn Lett 2010, 31, 651–666.
20. Kohonen, T.; Kaski, S.; Lagus, K.; Salojarvi, J.; Honkela, J.; Paatero, V.; Saarela, A. Self organization of a massive document collection. Neural Networks 2000, 11, 574–585.
21. Kuo, R.; Ho, L.M.; Hu, C.M. Cluster analysis in industrial market segmentation through artificial neural network. Comput Ind Eng 2002, 42, 391–399.
22. Kuo, R.; Ho, L.M.; Hu, C.M. Integration of self-organizing feature map and k-means algorithm for market segmentation. Comput Oper Res 2002, 29, 1475–1493.
23. Savage, M.; Burrows, R. The coming crisis of empirical sociology. Sociology 2007, 41, 885–899.
24. Savage, M.; Burrows, R. Some further reflections on the coming crisis of empirical sociology. Sociology 2009, 43, 762–772.
25. Mackey, M.C. Time's Arrow: The Origins of Thermodynamic Behavior. Springer Verlag, Germany; 1992.
26. Haken, H. Information and Self-Organization: A Macroscopic Approach to Complex Systems. Springer: Germany, 2006.
27. Christakis, N.A., Fowler, J.H. The spread of obesity in a large social network over 32 years. New Engl J Med 2007;357:370–379. DOI 10.1056/NEJMsa066082

DOI 10.1002/cplx