Journal of
open research software

Schimpf, C and Castellani, B 2020 COMPLEX-IT: A Case-Based Modelling and Scenario Simulation Platform for Social Inquiry. *Journal of Open Research Software,* 8: 25. DOI: https://doi.org/10.5334/jors.298

SOFTWARE METAPAPER

COMPLEX-IT: A Case-Based Modelling and Scenario Simulation Platform for Social Inquiry

Corey Schimpf* and Brian Castellani†

* University at Buffalo, Buffalo, US

† Department of Sociology, Durham University, UK

Corresponding author: Corey Schimpf (schimpf2@buffalo.edu)

COMPLEX-IT is a case-based, mixed-methods platform for applied social inquiry into complex data/systems, designed to increase non-expert access to the tools of computational social science (i.e., cluster analysis, artificial intelligence, data visualization, data forecasting, and scenario simulation). In particular, COMPLEX-IT aids applied social inquiry though a heavy emphasis on learning about the complex data/system under study, which it does by (a) identifying and forecasting major and minor clusters/trends; (b) visualizing their complex causality; and (c) simulating scenarios for potential interventions. COMPLEX-IT is accessible through the web or can be run locally and is powered by R and the Shiny web framework.

Keywords: Complex systems; computational modelling; case-based methods; social complexity theory; evaluation research; machine learning; data forecasting; data mining; data visualization

Funding statement: COMPLEX-IT was developed with support from the grants received by CECAN, namely, Economic and Social Research Council (grant numbers: ES/N012550/1 and ES/S000402/1). Funding was also received from Durham University via the ESRC pathways to impact grant.

(1) Overview

Introduction

I. The challenges of studying complex data

Many of the data sets and topics policy analysts, businesses, public agencies, and applied researchers struggle to understand are best described as case-based, complex, and applied. They are '*case-based*' insomuch as they are comprised of long profiles of information (e.g., variables, factors, conditions, etc.) about some set of cases, such as clients, patients, consumers, or communities. Given these cases and their profiles, a key goal is to find non-obvious patterns, such as how different cases cluster and why; or, for time-series data, how these clusters differ or change across time/space – for example, how a disease like COVID-19 spreads throughout communities or how a public transportation strategy impacts different communities during a policy cycle.

These datasets are '*complex*' insomuch as the case profiles are multi-dimensional, multi-level, dynamic, nonlinear, self-organizing, emergent, and geospatially and contextually (path) dependent [1, 2]. For example, data on community health is interdependent with household income, which is linked in complex nonlinear ways to job growth, education, air quality, and crime.

These datasets are '*applied*' insomuch as they are oriented around real-world issues, and because the goal of analysing them is to help influence, change or alter the course of something. This is particularly true in such areas as applied research, healthcare, education, public infrastructure, social services, and policy and program evaluation.

Working with these sorts of datasets creates several methodological challenges. First, most users are only trained in the conventional methods of statistics or qualitative inquiry. Second, even for those aware of the recent developments in computational modelling, data mining or big data analytics, these tools and techniques remain beyond everyday usage. Which, in turn, creates a third challenge, as users wanting to employ these new techniques often become over-reliant upon specialists that have no background or expertise in the topics they are being asked to model. Fourth, even for computational modelling experts, there is presently few packages available that integrates these techniques into a dedicated, seamless and visually intuitive platform, let alone ground a potentially diverse suite of techniques in a methodological framework sufficient to epistemologically stitch them together. Hence the purpose of COMPLEX-IT.

II. COMPLEX-IT: A case-based approach to social inquiry

Given the above methodological challenges, the purpose of COMPLEX-IT is to make computational modelling accessible to a wider non-technical and mostly applied

audience. COMPLEX-IT does two things. First, it improves the user-centeredness of these techniques. Second, it employs a case-based modelling approach [1–11].

A. User-Centred Approach

COMPLEX-IT increases the usability of computational modelling by distilling these methods into their essential features and streamlining their integration– which is accomplished in two ways: functionality and interface design. COMPLEX-IT's *functionality* is unique because it runs a specific suite of techniques that support case-based data exploration, modelling, forecasting and scenario simulation. In turn, COMPLEX-IT's *tab-driven interface* provides users a seamless, concise and visually intuitive platform. Also, advanced users can examine, download or modify COMPLEX-IT's algorithms, code, and outputs.

Currently (circa 2020), COMPLEX-IT's suite includes (1) k-means cluster analysis, (2) the Kohonen topographical neural net, (3) a series of data visualization techniques, (4) a machine intelligence algorithm for data forecasting, and (5) a tab for simulating and exploring future scenarios. The simulating scenarios tab is a major methodological advance, as it provides an alternative to agent-based modelling and microsimulation for exploring how to influence, change or alter the course of complex data/systems.

Case-Based Modelling

COMPLEX-IT is grounded in one of the major approaches for the study of complex social data: *case-based complexity* (CBC) [3–5]. Some of the most widely used CBC techniques include cluster analysis, machine intelligence [6], dynamic pattern synthesis [7] and Ragin's qualitative comparative analysis or QCA [8].

Regardless of the method used, CBC is anchored in five core epistemological arguments that deeply resonate with the majority of contemporary computational methods, as well as most users in the applied and public sectors. First, the case and its trajectory across time/space are the focus of study, not the individual variables or attributes of which it is comprised. Second, cases and their trajectories are treated as composites (profiles), comprised of an interdependent, interconnected sets of variables, factors or attributes. Third, the wider social contexts/systems in which cases are situated often needs to be considered. Fourth, in those instances where the relationships amongst a set of cases are important, network analysis of how cases interact and the structural patterns they form is key. And, finally, cases and their relationships and trajectories are the methodological equivalent of complex systems – that is, they are emergent, self-organizing, nonlinear, dynamic, etc – and therefore should be studied as such.

The specific CBC approach that COMPLEX-IT employs is called *case-based modelling*. The utility of case-based modelling is that it is a mixed-methods, computationally grounded approach to learning about and exploring complex social topics and datasets [9–11]. The methodological platform for case-based modelling is the *SACS toolkit* (sociology and complexity science toolkit), which provides a series of methodological steps and techniques (as well as a mathematical justification) for modelling complex systems in case-based terms [9].

For cross-sectional data, the purpose of COMPLEX-IT is to cluster multiple cases. A cross-sectional example would be clustering communities according to their levels of economic deprivation or community health and comparing that to rates of COVID-19. For temporal data, COMPLEX-IT clusters cases and their trajectories in the form of major and minor trends. By the term 'trajectories' we mean how a particular topic unfolds in different ways across time/space for diverse clusters of cases. The most common trajectories for the largest clusters are called major trends; and the least common are called minor trends. For example, a major COVID-19 trend was the high hospitalisation and mortality rate amongst the elderly. A minor trend would be that subset of young people with similarly high hospitalisation and mortality rates.

Also of importance for modelling temporal data is how autocorrelation and other such statistical phenomena play out in the analyses of clusters and their major and minor trends – which is why COMPLEX-IT is designed to visually and statistically data mine clusters and their trends for key global-temporal dynamics. For example, in the case of COVID-19 the spread of the disease in 2020 had a delay of about two weeks because of the complexities involved in its incubation, the potential of its spread by asymptomatic individuals, changes in social distancing, and the quality of testing. Given such issues, using COMPLEX-IT to compare and contrast clusters over time both visually and statistically to search for common global patterns (including delays) proves useful.

Finally, COMPLEX-IT also leverages its results to either predict novel cases or forecast future trends, as well as simulate different case-based scenarios. For example, given that Italy sheltered-in-place in 2020 several weeks before similar COVID-19 quarantines were launched throughout Europe, these data could be used to (a) cluster Italian provinces according to differences in their respective trends; and then (b) use this trained dataset to make short-term forecasts for matching authority districts in the UK or arrondissements in France. These data could also be explored, using the scenario simulation tab, to understand which types of social distancing strategies might be more useful than others and how alternative scenarios could prove beneficial.

Presently COMPLEX-IT cannot be used to model the relationships amongst a set of cases or the networks they form – such analyses are, however, under development with the addition of a tab for agent-based modelling tab with network support. (For an in-depth overview of COMPLEX-IT, including its mathematical foundation, see https://www.art-sciencefactory.com/cases.html).

Case-Based Scenario Simulation (COMPLEX-IT)

We call the integration of *case-based modeling* and *scenario simulation* Case-Based Scenario Simulation (CBSS). The purpose of CBSS is to create a simulated environment for users to visually and statistically explore different possible scenarios and outcomes for some set of case-based clusters/trends, which have been identified

earlier in the data analysis process [9, 19]. To do so, COMPLEX-IT draws inspiration from and builds upon the methodological traditions of (1) microsimulation and agent-based modelling; and explicitly integrates (2) case-based modelling (which encapsulates points 1–4 above) and (3) scenario analysis and planning approaches.

First, in terms of microsimulation and agent-based modelling [12, 13], CBSS models multiple clusters and their evolving trajectories across time/space, the latter of which leverages longitudinal clustering [9]. The difference, in other words, is that CBSS focuses on the clusters (not the cases) the user identifies during the earlier parts of their analyses. For example, in the Welsh multiple deprivation study provided on our support website, CBSS is used to explore different scenarios to help improve the cluster of communities with the highest deprivation.

For case-based modelling, COMPLEX-IT leverages k-means cluster analysis [14] as a user-driven way to identify major and minor clusters or (for time-series data) trends among a set of cases. The clusters or their trends identified by k-means are then corroborated and extended through the self-organizing map (SOM), an artificial neural network technique that preserves the topography of analysed data and which is commonly used in conjunction with k-means [9, 15]. From here, results are explored using a series of data visualization tools, in particular the SOM output grid (See **Figure 4** below).

As the final step, CBSS draws on scenario analysis and planning [16, 17], a broad collection of techniques for developing and evaluating scenarios effects on an entity of interest. Evaluating these scenarios, as in the case of how COVID-19 has spread amongst different communities, provides insight into how the entity might respond under uncertain circumstances and informs planning [18]. In COMPLEX-IT, the 'scenario simulation' component enables targeted exploration of how case-based clusters respond to various plausible scenarios they may encounter. For example, how urban areas with high density and poverty improve through a scenario involving contact tracing and wearing masks during the COVID-19 pandemic.

Thus, summarizing from above, scenario simulation in COMPLEX-IT offers users several strengths in the study of complex data/system. First, it allows users to explore how a cluster or its trend can be driven in a more desirable direction, by simulating different scenarios or interventions into its composite of causal conditions (i.e., its profile of factors, variables, measurements) and then running the model to see if the desired change took place. In other words, given its emphasis on learning, it can generate multiple and different models of complex data/systems that are flexible and evolving to the needs of the user. Generating multiple models is useful when there (1) is a high level of uncertainty around a cluster/trend of interest; (2) there are multiple plausible interventions that can take place; or (3) there are multiple events that could impactfully affect the cluster/trend.

Second, in addition to exploring what leads to a desired change, scenario simulations help the user learn about (1) how different clusters or their trends respond to plausible events; (2) how resilient they may be to changing classification in the state space; (3) what leads to undesirable change; and (4) the type and degree of intervention necessary to propel a trend toward another cluster/trend's profile.

Third, it offers the ability to analyse how different clusters or the entire complex system of study might react to various possible scenario changes or interventions in order to help users plan for the multiple contingencies and paths the cluster/trends and system face.

Finally, unlike agent-based modelling, COMPLEX-IT is always empirically dependent and driven, starting with the user's data. In other words, one must use data to employ the COMPLEX-IT approach.

Understanding the Limits of COMPLEX-IT

There are some important caveats to note about using COMPLEX-IT. First and foremost, the results from most modelling techniques, including COMPLEX-IT, rely on the assumption that a user's model is a reasonable approximation of its real-world counterpart. Care needs to be given when building a model to critically examine the ways it does and does not represent its real-world counterpart. For example, the cases being modelled may be poorly specified (e.g., lacking key case profile factors or case units may be too large or too small). To help surface when this disconnect may be happening, we recommend users employ regular validity checks to pull back from the analysis and consider whether their model is consistent with other data sources they have, past research findings, experience, or relevant theories [1].

Another issue is how the variable profiles COMPLEX-IT uses to cluster cases may be intercorrelated. Statistical terms used to discuss these issues include autocorrelation (which we discussed in the introduction regarding time-series data) and multicollinearity, which refers to a situation in which two or more explanatory variables in a causal model are highly related. This is a problem for all classification methods and is not unique to COMPLEX-IT. Still, users need to think through these issues. For more, see [11, 14].

We also strongly emphasize that COMPLEX-IT is a learning environment that requires users to be in direct and constant interaction with the results of their analyses and with their domain expertise, theoretical understanding, or theories of change, be they sitting implicitly in the background of their minds or formally outlined and defined. In other words, the goal of COMPLEX-IT is less about creating a statistically verified causal model (even though it can do that), as much as it is about exploring and learning new things about datasets that are case-based, complex and applied in focus, including how various interventions might unfold for a given policy or program and the larger complex system in which it is situated.

Finally, a few smaller limitations should also be noted. As an empirically driven analysis, COMPLEX-IT cannot depart far from the data it uses; and it cannot easily be used for forecasts or scenarios into the distant future, as there are too many contingencies and uncertainty to know where systems may be after a substantial time-gap.

Implementation and architecture

COMPLEX-IT (shown in **Figure 1**) was built with the R programming language and Shiny, a web-framework for R. It can be accessed through a browser for the server hosted version or downloaded and run with a localhost through the RStudio IDE. We use a Unified Modelling Language (UML) activity diagram to depict its architecture [20]. Activity diagrams present a series of activities a system progresses through within a use-session, including branching paths different sessions may take.

The activity diagram for COMPLEX-IT is displayed in **Figure 2**. Note the diagram displays the three core paths we've developed for COMPLEX-IT – the design of which was informed from feedback and requests from different user groups (particularly those in evaluation research). However, other activity flows can be employed. Within COMPLEX-IT each of the activities in **Figure 2** has a specific set of inputs, outputs and designated tasks. Each activity or tab is also accessible as a separate self-enclosed section of the web-app. Below we present each tab following the order of **Figure 2**. For more in-depth reviews of these tabs, including tutorials, please see our support website: https://www.art-sciencefactory.com/tutorials.html.

I. Data upload tab

The main purpose of this tab is to enable users to upload and shape their data for analysis. One example, which our support website provides, is a dataset of authority districts in Wales (the cases) and their scores on a list of multiple deprivation indices. Another example, which we also discuss in a tutorial video on our website, is trend data on COVID-19 for all N = 149 authority districts in the UK. Here the variables are daily cumulative case counts gathered from the Public Health England website.

Any uploaded data should be in *CSV* format and follow the structure of rows as cases and columns as the profile components of the cases. Only data with numeric values can be analysed in COMPLEX-IT. Upon uploading data, an adjustable preview of the rows and columns will appear at the bottom of the dashboard. Finally, users can modify which elements are included in cases' profiles before starting analysis by sub-setting the uploaded case data.

There are two brief notes to highlight about data pre-processing. First, there is an option to standardize (centre and scale) your data on the SOM analysis tab. We encourage users to consider standardization when their cases profile variables are on different scales (e.g., age and income) as variables with a larger scale—in contrast with those on a smaller scale—will have a dominating influence on cluster results. We recommend using the scale function built into R itself to centre and scale data before uploading it into COMPLEX-IT when users find themselves in these circumstances. Second, COMPLEX-IT will discard cases with missing numeric data as these will cause problems for the clustering algorithms. Please assess your data beforehand and consider data imputation techniques when missing data is a challenge. For more, see [11, 14, 15].

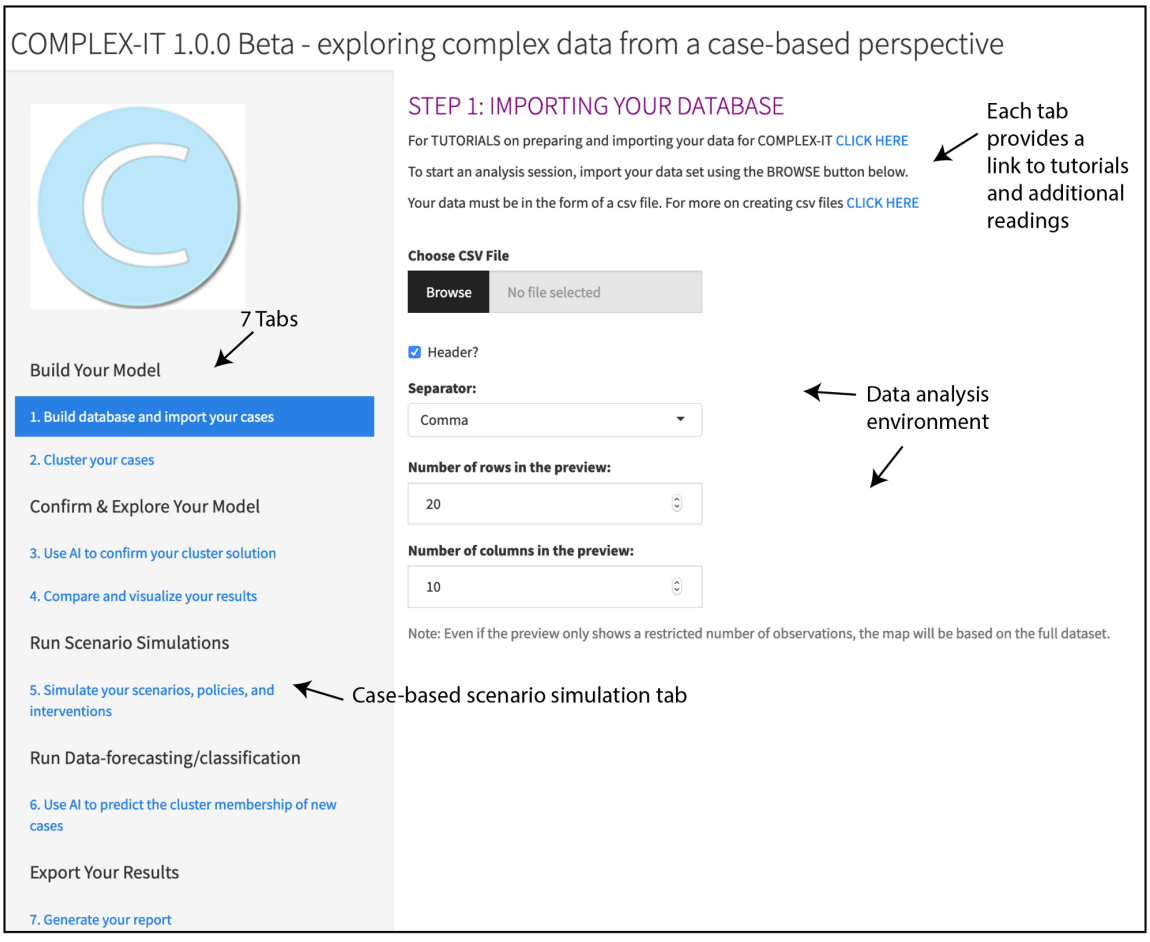

Figure 1: COMPLEX-IT Interface.

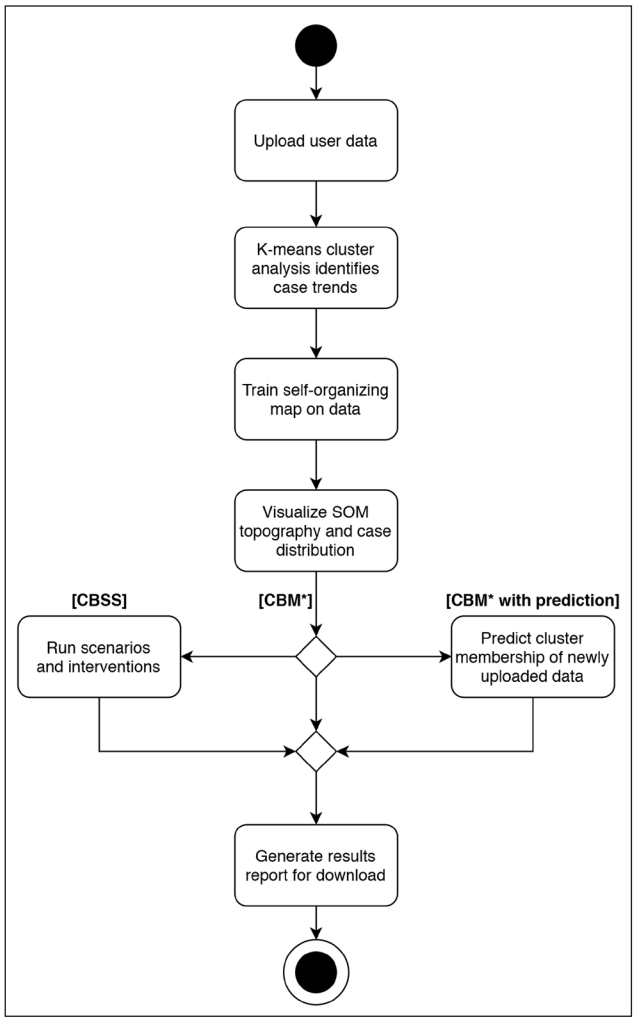

Figure 2: UML Activity Diagram for COMPLEX-IT. * Case-based modelling.

II. Cluster your cases tab

The main purpose of this tab is to enable users to group cases together in order to identify major and minor clusters or trends. For example, for our cross-sectional Wales dataset, we sought to cluster authority districts based on their multiple deprivation score; and for the spread of COVID-19 during 2020 we sought to cluster regions of the UK based on changes in their daily cumulative case counts.

To create these clusters, this tab employs the k-means clustering algorithm because it requires users to apply their experience and knowledge of their topic of study to select an appropriate number of groups or clusters and to evaluate the validity and empirical sensibility of the output (See **Figure 3**). For example, in terms of Welsh multiple deprivation, one could work with public health experts to identify how many clusters might exist in the dataset. And, for COVID-19 one could consult with local trusts and councils.

Overview of K-means and COMPLEX-IT output

K-means (which is a type of semi-supervised learning) operates by separating cases into groups by minimizing the within group differences between a cluster's associated cases [12]. K-means then iteratively moves cases between groups to achieve this goal. A final cluster is represented by a centroid, which contains the average values for each element of the case profiles in each cluster. Thus, distributions across clusters ideally are tightly packed around the centroid and unique from other clusters in the set.

Upon running the k-means algorithm, COMPLEX-IT will present the resulting cluster profiles and size of the clusters. Additionally, the user can request evaluations on the quality of the clustering: *pseudo f* and *silhouette plots*, two commonly used metrics for evaluating clusters. *Pseudo f* is an overall measure of how tightly cases are grouped within clusters and how separate or non-overlapping clusters are, with higher values indicating better performance. The *silhouette plot* allows for visual and quantitative inspection of cases fit within their clusters, see **Figure 3**. Through inspection of the clusters vis-à-vis domain knowledge or relevant theory and these quality measures, users can identify the best arrangement for the major and minor trends in their set of cases.

III. SOM training tab

The main purpose of this tab and the following tab is to enable the user to employ the self-organizing map neural net (SOM) – which is a form of artificial intelligence – to do three things: corroborate the k-means solution, or explore the k-means solution in greater depth, or run the SOM as its own clustering solution, without running k-means. Note, however, both the SOM and k-means need

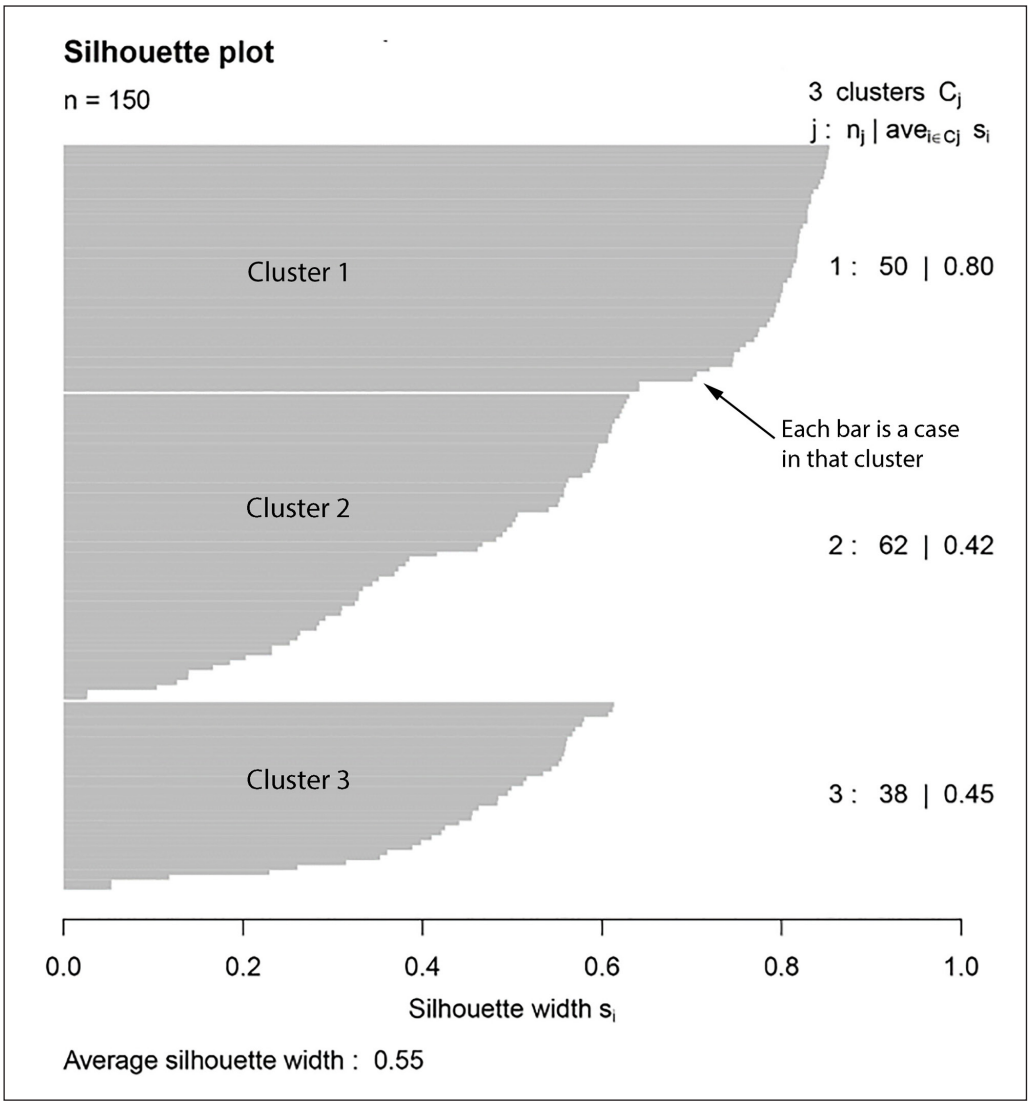

Figure 3: Silhouette plot. Each cluster is plotted as a horizontal bar-plot, with a bar for each case. Bar values span the range of –1 to 1, with near 1 being a strong fit and zero or less being a very poor fit. The 'average width' or fit within each cluster is displayed on the right after the cluster number and size; and the 'overall average fit' is found at the bottom.

to be run to use the scenario simulation tab and the prediction/data forecasting tab. For example, in the case of multiple deprivation in Wales, the SOM can be used to corroborate the k-means solution. And, for COVID-19, it could be run without the k-means solution in order to generate a more granular understanding of the sub-clustering of authority districts in the UK given the wide variation in time-series trends.

Overview of the SOM and its output

The SOM is a topographical neural network designed for unsupervised learning (i.e., machine intelligence), data visualization, and clustering [11, 15, 21]. This value of this approach is that, while users are required to set up some of the initial parameters, no set number of clusters are prechosen. Instead, the SOM is left to identify the key clusters and, for temporal data, the major and minor trends.

For example, while k-means requires users to look for a specific number of clusters amongst a dataset of COVID-19 cases, the SOM identifies the number of clusters that best fit the data based on the algorithm. It is because of this difference in approach that experts in the field recommend the SOM solution compared to the k-means solution to corroborate results or suggest further analysis [11, 15].

We chose SOM over other neural network techniques because it is a well-established approach with unique features for supporting human analysis of results [11, 15]. Unlike other artificial neural network techniques, the SOM classification model can be directly visualized and analysed to understand how cases were assigned to a neuron (cluster) outcome – which is the focus of the following tab. The classification model is directly interpretable because the SOM projects high-dimensional (multiple variable) input data onto a 2-dimensional grid solution (See **Figure 4** below) that can be visualized in a variety of ways – which we discuss in the SOM analysis section. Thus, the SOM provides users, including those less familiar with machine learning techniques, a more direct way to compare their k-means and SOM results.

Options for running the SOM training Tab

The options on this tab concern SOM algorithm setup, including setting the size of the grid, weights initialization, the number of algorithm iterations, data scaling, and a seed for preserving the initialization and the learning rate. After training the SOM, ANOVA is run to provide information on which (if any) of the profile elements differed significantly across the neurons. This lends insight into what may be distinguishing factors across the neurons. Additionally, two quality measures are displayed, the *quantization error* and the *topographical error*. The first indicates how much, on average, cases diverge from their assigned neuron; and the second addresses the rate at which surrounding neurons are a good fit for a neuron cases across the grid. Preferably these have lower values close to zero, as they are measures of error.

IV. SOM analysis tab

The main purpose of this tab is to allow for an immediately intuitive and visual inspection of the SOM solution, onto which the k-means solution is also projected for those instances in which both techniques are being used, as in the case of corroboration, or for later analyses using the tabs for scenario simulation or prediction and forecasting. In addition, the tab for generating your report provides details statistical information, including Excel databases that can be immediately explored.

In the case of deprivation in Wales or COVID-19 throughout the UK, all of the authority districts could be visually explored on this tab to see their profiles, bar-plots, the cases that belong to each quadrant of the map and also their k-means solution, as well as other information. When working with stakeholders, they report that the immediacy of this information is useful. It also proves useful in that, if they want to change things, they can go back to the clustering or SOM training tab and run things again and then immediately see the different results in this tab.

Understanding the SOM Map

The central object of study for this tab is the SOM map, shown in **Figures 4** and **5**. As mentioned earlier, the SOM 'maps' complex data by creating a grid of $n \times n$ neurons. On this map, each neuron has a set of weights equal to the number of configurational variables in the case profile (i.e., causal conditions, factors, measures, etc.), which makes each neuron somewhat similar to a k-means cluster centroid. For example, looking at **Figures 4** and **5**, one sees a 5×5 SOM Map, with a total of 25 neurons (i.e., cluster solutions). We happen to call these neurons *quadrants*, which came from our work with various users, who favoured this terminology.

One also sees in **Figure 4** that some of the quadrants do not have a bar-plot configuration. The same with **Figure 5**: some of the quadrants do not have cases. The reason is that these quadrants did not constitute viable cluster solutions for the data. And it is because of these "empty" quadrants, in part, that the SOM is referred to as *unsupervised learning*, as it lets a dataset of cases settle (map) onto the 'best' cluster solution across all quadrants, with cases iteratively finding the 'right' quadrant on which to reside. Unsupervised learning takes place by the SOM setting default weights for all the quadrants and then associating cases with the quadrants they most closely resemble, usually through a distance measure. The weights of the quadrants and their neighbours are then updated based on an adjustable learning factor, which also leads to similar cases 'settling' into comparable regions of the map. The degree of 'learning' for each quadrant (and its capacity to attract similar cases in its region) decreases over time. The SOM is assessed, as mentioned earlier, by the two validity measures *quantization error* and the *topographical error*.

In short, the SOM's 'topographical' mapping of the data leads to quadrants in proximity having similar weights for their configuration of variables (i.e., factors, causal conditions, measures, etc.) and those further away having

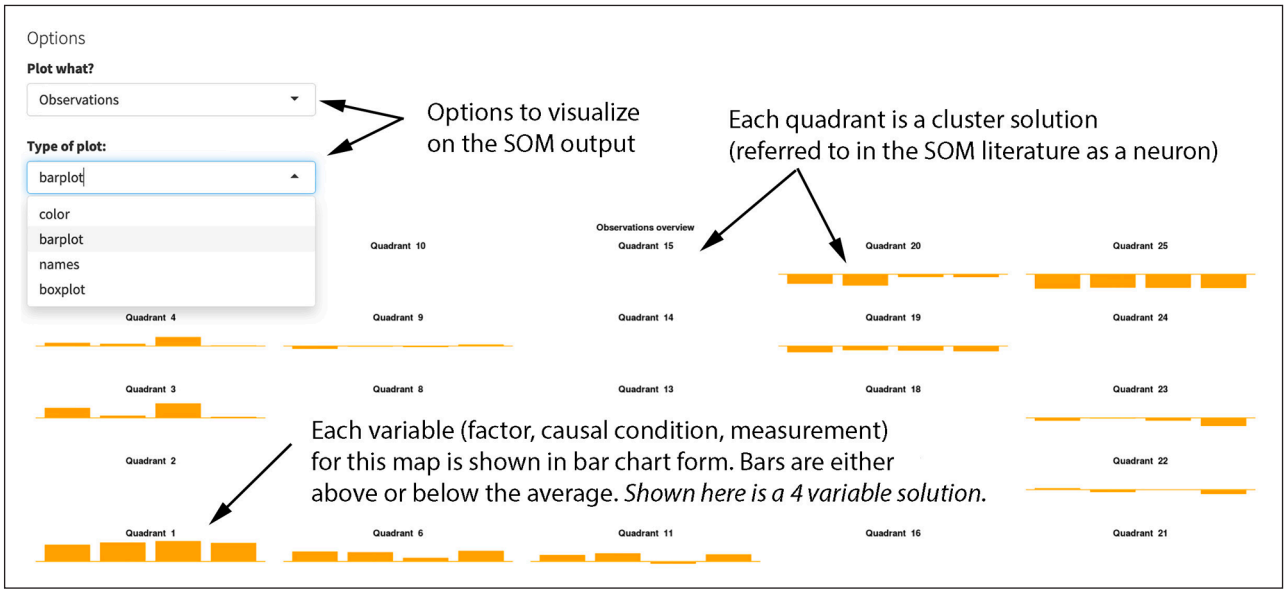

Figure 4: SOM Bar-plot. Bars represent the average value for case elements in each quadrant.

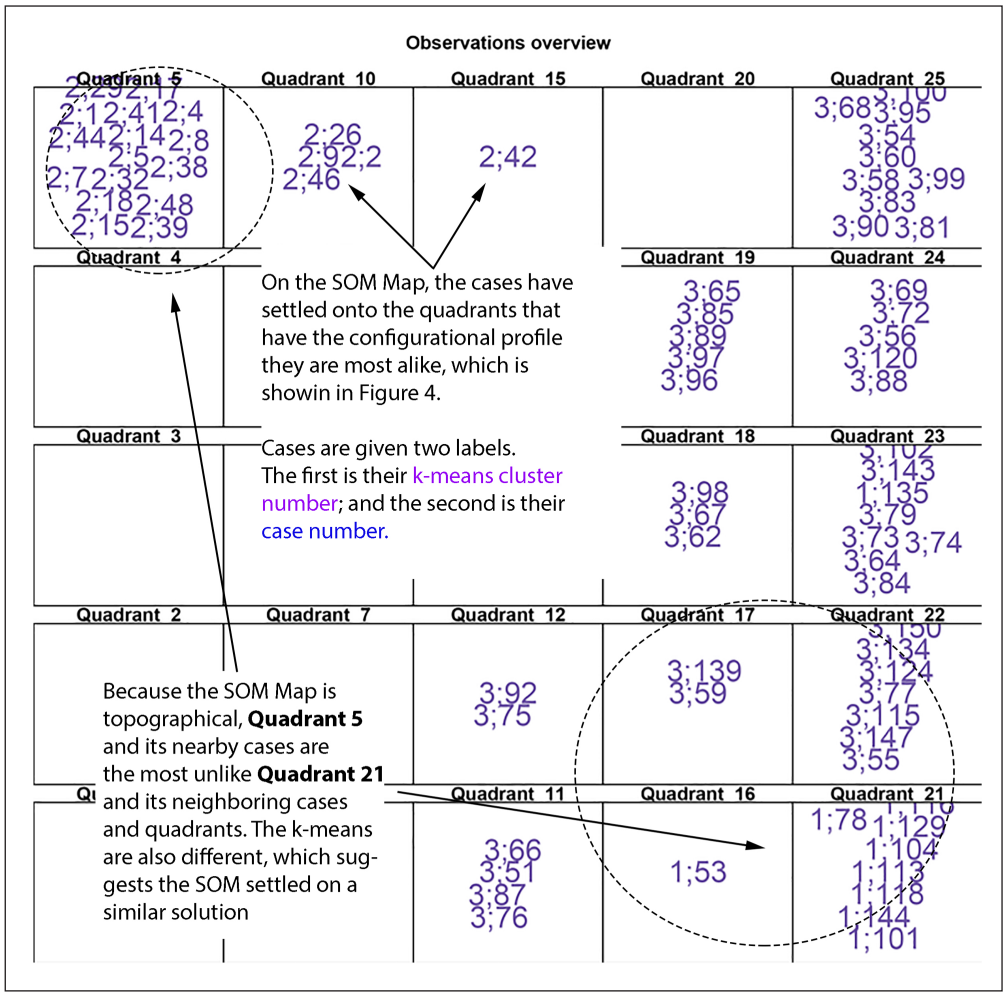

Figure 5: SOM Names plot. Note, for each plotted case cluster ID is first and case ID is second.

progressively larger differences in weights. In other words, quadrants (and hence cases) that are near each other on the map are more alike configurationally than those further apart. Hence why this approach is called the self-organizing topographical map.

Options for viewing the SOM map Tab

As suggested above, the strength of this tab for the SOM map is the variety of options it provides for visualizing and analysing the SOM's results. They are broken down into two types: prototypes (i.e., the configurational factors/variables in one's study) and observations (i.e., cases in the dataset).

Figure 4, for example, is a bar-plot of a study exploring four factors. The bar-plot displays the entire SOM grid and each quadrant will contain a series of bars or no bars if no cases were assigned to that particular quadrant. Each bar in a quadrant represents one of the elements of the cases' profiles and should be read from left to right. Similar to k-means, the bars for each quadrant represent the average values for cases within it. Bars are centred on the global mean for any given case variable; therefore, if a bar is only a line within the quadrant, it is at the global mean, whereas bars higher or lower than the centre line indicate it is higher or lower than the global mean, respectively. In this way the bar-plot provides a view into the major and minor trends for the cases through quadrant profiles. As some quadrants may have similar profiles, full numeric data for each quadrant can be retrieved from the tab for generating your report. Quadrant profile data can then be used to group quadrants based on similarity and thereby simplify the set of major and minor trends.

Corroborating the SOM and k-means solutions

For some analysis, users may wish to compare the SOM solution to the k-means solution or even use one to corroborate the other. When corroborating, we recommend the two-step procedure advocated by [15], which found that a combination of k-means and SOM is often more useful (particularly for reducing misclassification) than the conventional approach of using just one technique or combining hierarchical clustering and k-means.

In terms of statistical validation, 'goodness of fit' for k-means is measured by *pseudo f* and *silhouette plots*; and for the SOM it is *quantization error* and *topographical error*. The first step in corroboration is to make sure both solutions fit the data.

Next is visual inspection of the SOM map. Using the names plot, users can explore how the k-means clusters are distributed across the map and the SOM cluster profiles they are associated with. Plotted cases follow the convention of displaying cluster ID and case

ID overlaid on the SOM map. Two things need to be considered when making such comparisons. First, if one runs a large SOM map or chooses a sizable k-means solution, the resulting map will show a more fine-grain distribution of cases. In turn, smaller maps and k-means solutions will give the impression of things fitting better. The real comparison comes from using the bar-plot option to look for similarities in profiles: one is looking to see if the k-means profile for a case is similar to the SOM profile for the quadrant it is in. Second, visual inspection, while useful, has its limits. The report generation tab creates Excel files with the cluster and quadrant IDs for all cases. These files can be statistically explored further, which is particularly useful when using large datasets.

V. Simulating scenario/intervention tab

Simulation is a powerful exploration tool, particularly for users in the public sector, policy evaluation, or applied fields of study, where having a low-risk environment to understand how, when, and where effective interventions can be made for some population or complex system of study is paramount. In the case of multiple deprivation in Wales or COVID-19 in the UK, for example, the scenario simulation tab allows users to explore different community-based interventions to see if they produce the types of results expected based on the users informal or formal theory of change or policy/program goals. For Welsh deprivation, scenarios could involve policy interventions to improve highly deprived communities. Or for COVID-19 time-series data, scenarios could involve applying a successful policy from one community to different communities.

Simulation, however, comes with its own challenges. One challenge is staying close to the data of study; another is the need to remain focused on the population of cases (as opposed to variables) in a study; and, finally, there is the need to build a model that is simple and yet complex enough to extract meaningful information. We would also remind users to keep in mind the caveats from the introduction about using COMPLEX-IT to model complex systems, as these also apply to scenario simulation. Hence the purpose of scenario simulation. Given we reviewed the purpose and aims of scenario simulations in the introduction, we will focus here, instead, on how the tab works.

The goal of scenario simulation

The main purpose of this tab is to enable users to simulate different scenarios that the case trends in their dataset could experience, as well as identifying any associated contingencies or responses they may express or failures to change. Here we define 'interventions' as a particular type of simulated (but potentially real) scenario reflecting some proactive attempt to change one or more of the configurational factors for a cluster or trend (i.e., variables, measurements, causal conditions). Examples of interventions could be wider changes in the settings or systems in which the clusters are situated (i.e., economic or environmental changes); or they could be specific policy changes (i.e., improve schooling or access to health care) or changes within the cluster by its own cases (i.e., a community-level shift in thinking or behaviour). Other scenarios a user may wish to explore include reactive events originating from environmental forces or outside of proactive action, such as out-migration in a community. Better understanding from scenario simulations can also lead to more informed planning or strategic action, which may prove useful for areas like policy analysis and program evaluation.

Understanding the scenario simulation map

To begin, scenario simulation combines results from the previously trained SOM and k-means analysis, presenting the same SOM map used in the SOM analysis tab. However, unlike the previous SOM map, which plots all the cases in the study, this map plots the k-means clusters identified previously.

Scenario simulation takes this focus because the k-means solution offers a more concise user-identified summary. For example, if one had a study with N = 100 cases, simulating interventions into each and every one of them may prove unmanageably complex and time consuming. Still, given that the SOM map from the SOM analysis tab is preserved, the cases for any given cluster and their respective differences can always be re-examined. In other words, while exploring the scenario map, one can also explore the prior SOM Map to gain a more granular understanding of how interventions or changes made on the scenario map might differentially "play out" at the case level.

Running a simulation

By pressing the *Model Setup* and *Run Clusters* buttons, the scenario tab will be initialized, as shown in **Figure 6**. Below the SOM plot an editable table that allows the user to update one or more elements of the k-means case trends. Different scenarios can be explored by changing the relevant elements in the table for a given scenario and pressing *Run Clusters*. The updated case trend will be re-examined against the SOM quadrants and the impact of the scenario, if any, will be shown through which quadrant it is now most closely associated with. The SOM bar-plot is also plotted here as a reference for the quadrant profiles.

Exact estimation of the change a scenario causes a case trend may not be possible. Therefore, the scenario tab also offers sensitivity analysis to account for uncertainty in the efficacy of scenario influencing a case trend. The user can specify how much a change to one or more element in a case trend may deviate by, from zero to one hundred percent of the submitted change, both in the positive and negative direction. This can be accessed by selecting the case trend/cluster to test and pressing the *Sensitivity* button on the left panel. After inputting the deviation for changed element(s), a Monte Carlo simulation will randomly sample from across this range, testing and providing a summary of which quadrants the case trend would be associated with – which allows a key insight into how 'sensitive' the effect of the scenario is to projected deviations.

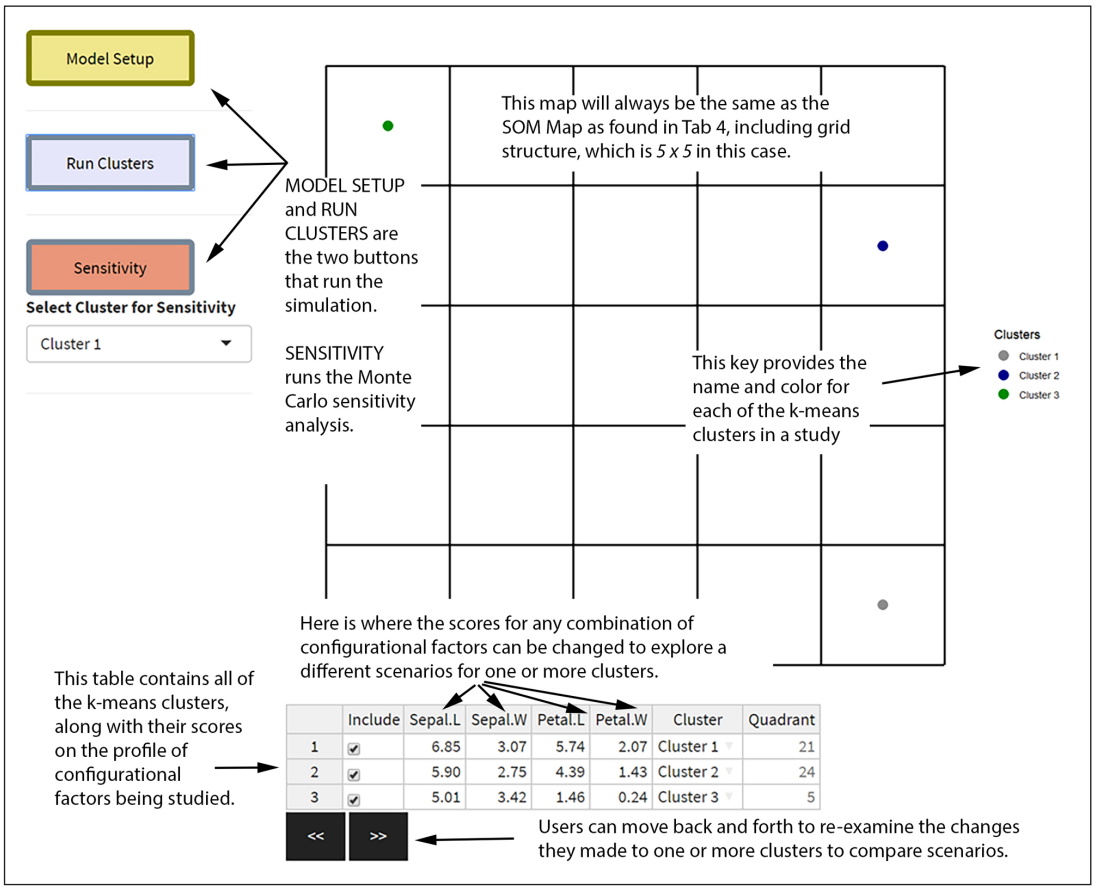

Figure 6: Scenario and Intervention tab main display and controls.

VI. Prediction and forecasting tab

This tab was created to help users involved in fields where they are regularly asked to make predictions for new or different cases or to forecast existing or new trends. Examples for cross-sectional data include identifying "at risk" urban communities due to air pollution deprivation. In terms of time-series, an example would be short-term forecasts on how an infectious disease like COVID-19 might trend for different countries based on comparing them to Italy (as we mentioned earlier).

Like the SOM analysis and scenario tabs, this tab uses the trained SOM as its map. However, the focus here is back to the cases and the SOM map quadrants. In other words, the goal is to predict where a new case or cases belong on the prior map, based on the SOM solution arrived at in the initial dataset studied.

The process is rather straightforward. Similar to the Data Upload tab, data can be uploaded to this tab. It is critical, however, that the uploaded data used in this tab have the same data columns as the data originally uploaded in COMPLEX-IT for any given session. Otherwise COMPLEX-IT will reject the new data and display a warning, as the SOM cannot make predictions on variables that were not part of its training set. After classifying the new dataset, a table will display the cases as well as their best-fitting and second best-fitting quadrant; the bar-plot from the SOM is also displayed as a reference for the quadrant profiles.

VII. Generate report tab

The main purpose of this tab is to generate exportable results for the tabs a user has employed. The generate report tab can be accessed any time to export detailed results from the k-means and SOM analysis as well as the scenario and predict tab. Results will only be exported for those tabs the user has employed in the current session; so long as at least one tab has been used it is possible to export results. Exported results will reflect the most recent analysis, simulated scenario, or prediction performed.

More comprehensively, k-means results include the cluster profiles, *pseudo f*, cluster ID for all cases and the silhouette. SOM results include the parameters of the SOM (e.g., learning rate), the ANOVA results and quality measures, the quadrant profiles, quadrant IDs for cases as well as the bar-plot and boxplot graph. Scenario results include the intervention tested and sensitivity analysis results while the prediction results include the new data and their assigned SOM quadrant. Exportable results are intended to allow the user to share or disseminate findings or conduct further analysis.

Quality control

COMPLEX-IT has been developed through an agile process over the past three years, with team members regularly meeting, planning, and implementing new features which were then reviewed and revised as the platform evolved. During this time, the development team tested new

features as they were introduced to ensure they performed as expected. These were subject to regular full platform reviews to check that all components were properly integrated. Additionally, the primary team members regularly analysed new datasets through COMPLEX-IT. Any technical bugs, interface difficulties or other challenges were reported to the development team and added to the ongoing meetings to resolve.

(2) Availability

Operating system

The hosted version of COMPLEX-IT is accessed through a browser and is compatible with most modern web browsers. The downloadable version of COMPLEX-IT is compatible with Windows 7 or MacOS 10.7 or later versions of Windows or Mac. It is also available on Debian, SUSE, Ubuntu and Redhat Linux distributions.

Programming language

R 4.0.2 or above.

Additional system requirements

No additional requirements for the hosted version of COMPLEX-IT. If users want to run locally, it will be necessary to installed R and RStudio, which require a minimum of 256 megabytes of RAM.

Dependencies

All dependencies are packages or frameworks for R.
cluster 2.1.0 or higher
ggplot2 3.3.2 or higher
rhandsontable 0.3.7 or higher
shiny 1.5.0 or higher
shinyThemes 1.1.2 or higher
SOMbrero 1.3.0 or higher
zip 2.0.4 or higher

List of contributors

Brian Castellani project lead, researcher, and developer
Corey Schimpf lead developer and researcher
Mike Ball server administrator and developer
Peter Barbrook-Johnson project mentor and researcher
Nigel Gilbert project mentor

Software location

Code repository

Name: GitHub
Identifier: https://github.com/Cschimpf/Complex-It
Licence: MIT
Date published: 31/07/20

Language

English

(3) Reuse potential

Case based modelling and scenario simulations embodied in COMPLEX-IT, inquiry keeping in mind its strengths and limitations, hold great potential for several groups looking to engage in social inquiry. Most prominently, applied researchers and analysts in areas such as healthcare, education, public infrastructure, social services, and policy and program evaluation must confront understanding and affecting open systems e.g, see [22], which is supported through identifying different complex systems, trajectories when time-series data is available, as well as the impact of various interventions into the modelled system. Social scientists interested or already involved in the study of complex data/systems likewise can leverage COMPLEX-IT to expand our fundamental understanding of these systems. Finally, COMPLEX-IT may also be used as exploratory datamining tool for new or understudied topics.

Acknowledgements

The authors wish to thank Centre for the Evaluation of Complexity Across the Nexus (CECAN) for their financial and intellectual support, as well as Durham University for its financial support. Finally, we would like to thank Carl Dister for his development support in the early stages of COMPLEX-IT.

Competing Interests

The authors have no competing interests to declare.

References

1. **Castellani, B** and **Hafferty, F** 2009 Sociology and Complexity Science: A New Field of Inquiry. Berlin: Springer.DOI:https://doi.org/10.1007/978-3-540-88462-0
2. **Cilliers, P** 1998 Complexity & Postmodernism: Understanding Complex Systems. London: Routledge.
3. **Byrne, D** and **Ragin, C C** (Eds.) 2009 Case-Based Methods. Thousand Oaks, CA: Sage.
4. **Byrne, D** and **Callaghan, G** 2013 Complexity theory and the social sciences: The state of the art. London: Routledge. DOI: https://doi.org/10.4324/9780203519585
5. **Haynes, P** 2018 Dynamic Pattern Synthesis for Management, Business and Economics. Chichester, UK: White Horse Books.
6. **Castellani, B** and **Rajaram, R** 2019 Data Mining Big Data: A Complex Critical Introduction. *Part of the forthcoming Sage Quantitative Methods Kit*. Sage.
7. **Haynes, P** 2017 Social synthesis: Finding dynamic patterns in complex social systems. London: Routledge. DOI: https://doi.org/10.4324/9781315458533
8. **Ragin, C C** 2014 The comparative method: Moving beyond qualitative and quantitative strategies. University of California Press. DOI: https://doi.org/10.1525/9780520957350
9. **Castellani, B** and **Rajaram, R** 2012 Case-based modelling and the SACS Toolkit: a mathematical outline. *Comput Math Organ Theory*, 18: 153–174. DOI: https://doi.org/10.1007/s10588-012-9114-1
10. **Rajaram, R** and **Castellani, B** 2012 Modelling complex systems macroscopically: Case/agent-based modelling, synergetics, and the continuity equation. *Complexity*, 18: 8–17. DOI: https://doi.org/10.1002/cplx.21412

11. **Castellani, B, Rajaram, R, Gunn, J** and **Griffiths, F** 2016 Cases, clusters, densities: Modelling the nonlinear dynamics of complex health trajectories. *Complexity*, 21: 160–180. DOI: https://doi.org/10.1002/cplx.21728
12. **Mitton, L, Sutherland, H** and **Weeks, M** (Eds.) 2000 Microsimulation Modelling for Policy Analysis: Challenges and Innovations. Cambridge, UK: Cambridge University Press.
13. **Gilbert, N** and **Troitzsch, K** 2005 Simulation for the Social Scientist. New York, NY: Open University Press.
14. **Jain, A K** 2010 Data clustering: 50 years beyond k-means. *Pattern Recognition Letters*, 31: 651–666. DOI: https://doi.org/10.1016/j.patrec.2009.09.011
15. **Kuo, R J, Ho, L M** and **Hu, C M** 2002 Integration of self-organizing feature map and K-means algorithm for market segmentation. *Computers & Operations Research*, 29: 1475–1493. DOI: https://doi.org/10.1016/S0305-0548(01)00043-0
16. **Chermack, T** 2005 Studying scenario planning: Theory, research suggestions, and hypotheses. *Technological Forecasting and Social Change*, 72: 59–73. DOI: https://doi.org/10.1016/S0040-1625(03)00137-9
17. **Schwartz, P** 1991 The Art of the Long View: Planning for the Future in an Uncertain World. New York, NY: Doubleday.
18. **Börjeson, L, Höjer, M, Dreborg, K-H, Ekvall, T** and **Finnveden, G** 2006 Scenario types and techniques: Towards a user's guide. *Futures*, 38: 723–739. DOI: https://doi.org/10.1016/j.futures.2005.12.002
19. **Castellani, B, Barbrook-Johnson, P** and **Schimpf, C** 2019 Case-based methods and agent-based modelling: bridging the divide to leverage their combined strengths. *International Journal of Social Research Methodology*, 22: 403–416. DOI: https://doi.org/10.1080/13645579.2018.1563972
20. **Booch, G, Rumbaugh, J** and **Jacobson, I** 2005 The Unified Modelling Language User Guide, 2nd ed. Upper Sadle River, NJ: Pearson Education, Inc.
21. **Huang, T S, Kohonen, T** and **Schroeder, M R** (Eds.) 2000 Self-Organizing Maps, 3rd ed. Berlin: Springer. DOI: https://doi.org/10.1007/978-3-642-56927-2
22. **Barbrook-Johnson, P, Schimpf, C** and **Castellani, B** 2019 Reflections On the Use of Complexity-Appropriate Computational Modeling for Public Policy Evaluation in the UK. *Journal on Policy and Complex Systems*, 5(1): 55–70. DOI: https://doi.org/10.18278/jpcs.5.1.4

How to cite this article: Schimpf, C and Castellani, B 2020 COMPLEX-IT: A Case-Based Modelling and Scenario Simulation Platform for Social Inquiry. *Journal of Open Research Software,* 8: 25. DOI: https://doi.org/10.5334/jors.298

Submitted: 17 September 2019 **Accepted:** 17 September 2020 **Published:** 07 October 2020

]u[*Journal of Open Research Software* is a peer-reviewed open access journal published by Ubiquity Press.

OPEN ACCESS