This article has been accepted for publication in IEEE Transactions on Visualization and Computer Graphics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2022.3209464

Development and Evaluation of Two Approaches of Visual Sensitivity Analysis to Support Epidemiological Modeling

Erik Rydow, Rita Borgo, Hui Fang, Thomas Torsney-Weir, Ben Swallow, Thibaud Porphyre, Cagatay Turkay, and Min Chen, *Member, IEEE*

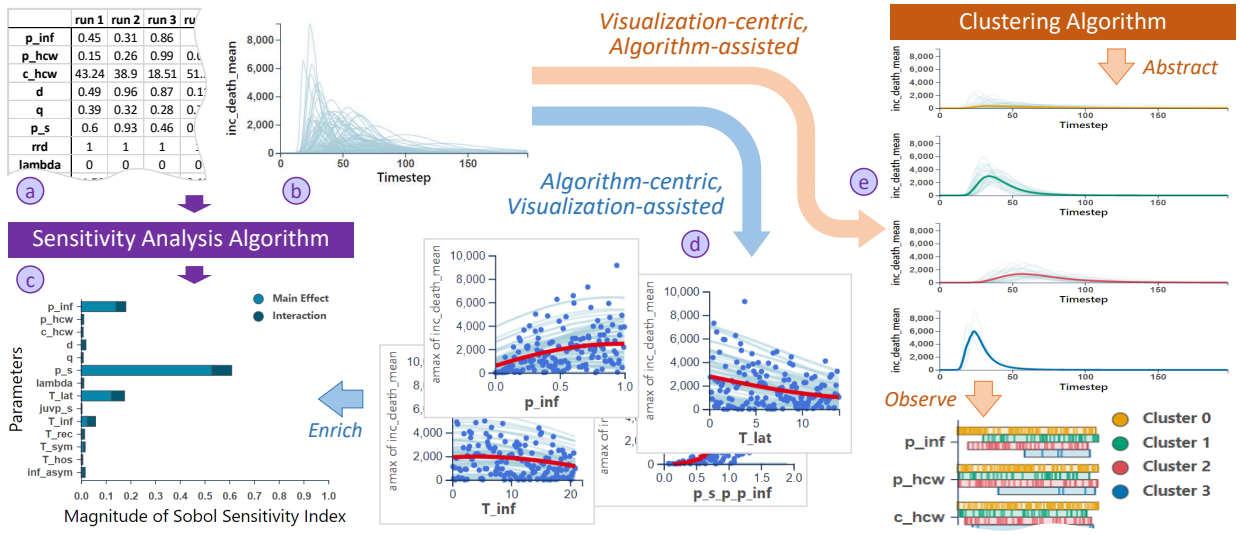

Fig. 1. An algorithmic approach to sensitivity analysis typically takes the data of ensemble runs, such as the parameter sets (a) and outputs (b), and estimates the sensitivity of each parameter (c). This algorithmic-centric approach can be assisted by visualization (d) that enriches the basis numerical measures, and be complemented by a visualization-centric and algorithm-assisted approach (e).

Abstract— Computational modeling is a commonly used technology in many scientific disciplines and has played a noticeable role in combating the COVID-19 pandemic. Modeling scientists conduct *sensitivity analysis* frequently to observe and monitor the behavior of a model during its development and deployment. The traditional algorithmic ranking of sensitivity of different parameters usually does not provide modeling scientists with sufficient information to understand the interactions between different parameters and model outputs, while modeling scientists need to observe a large number of model runs in order to gain actionable information for parameter optimization. To address the above challenge, we developed and compared two visual analytics approaches, namely: *algorithm-centric and visualization-assisted*, and *visualization-centric and algorithm-assisted*. We evaluated the two approaches based on a structured analysis of different tasks in visual sensitivity analysis as well as the feedback of domain experts. While the work was carried out in the context of epidemiological modeling, the two approaches developed in this work are directly applicable to a variety of modeling processes featuring time series outputs, and can be extended to work with models with other types of outputs.

Index Terms— Sensitivity analysis, Ensemble visualization, COVID-19, Epidemiological Modeling, Epidemiology.

1 Introduction

Epidemiological models have been playing an important role in combating the COVID-19 pandemic [2, 22, 30]. When the pandemic unfolded, modeling scientists had to develop models rapidly and improve them iteratively. A good portion of R&D activities in COVID-19 modeling have been to study how a model and its parameters have behaved at different stages of the pandemic and whether it would be helpful to incorporate new information into the model by modifying its parameter set (and less commonly its structure). These activities typically involve running a model with different parameter sets (*ensemble simulation*), estimating the amount of uncertainty in model outputs (*uncertainty quantification*), and establishing the relationships between inputs and outputs (*sensitivity analysis*). This work was conducted in collaboration with a team of epidemiologists and modeling scientists (hereafter referred to as "domain experts" for short), focusing on combined uses of algorithms and visualization in sensitivity analysis (SA).

Most modeling scientists are familiar with algorithm-based SA [44], where algorithms are used to analyze the impact of parameter inputs on the model outputs, and assign a scalar value of sensitivity to each parameter of the model as illustrated in Fig. 1(a-c). Such an analytical algorithm can measure and order the sensitivity of different parameters fast and consistently, as well as identify unexpected sensitivity or insensitivity of a parameter. However, the sensitivity values are not ground truth since different sensitivity algorithms normally yield

- *Erik Rydow and Min Chen are with Department of Engineering Science, University of Oxford, UK. E-mail: {erik.rydow, min.chen}@eng.ox.ac.uk.*
- *Rita Borgo is with Department of Informatics, Kings College London, UK. E-mail: rita.borgo@kcl.ac.uk.*
- *Hui Fang is with Department of Computer Science, Loughborough University, UK. E-mail: h.fang@lboro.ac.uk.*
- *Thomas Torsney-Weir is with Zentrum für Virtual Reality und Visualisierung Forschungs-GmbH (VRVis), Austria. E-mail: torsney-weir@vrvis.at.*
- *Ben Swallow is with School of Mathematics and Statistics, University of Glasgow. E-mail: ben.swallow@glasgow.ac.uk.*
- *Thibaud Porphyre is with Laboratoire de Biométrie et Biologie Evolutive (LBBE), VetAgro Sup, France. E-mail: thibaud.porphyre@vetagro-sup.fr.*
- *Cagatay Turkay is with Centre for Interdisciplinary Methodologies, University of Warwick, UK. E-mail: cagatay.turkay@warwick.ac.uk.*

Manuscript received xx xxx. 201x; accepted xx xxx. 201x. Date of Publication xx xxx. 201x; date of current version xx xxx. 201x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: xx.xxxx/TVCG.201x.xxxxxxx

different values. The information conveyed by these values is usually not sufficient to capture all meaningful variations of complex model outputs or in informing any action to improve a model by modifying the specification of its parameters.

To help domain experts address the shortcomings of the algorithm-only approach, we introduce data visualization to enable domain experts to gain more actionable information by observing more detailed interactions between the parameter inputs and model outputs. Drawing from existing techniques for ensemble, uncertainty, and sensitivity visualization, mostly in applications of computational fluid dynamics [60], weather simulation [5], and industrial modeling [35], we formulate two visual analytics approaches to support SA in epidemiological modeling. Our months long collaboration as an interdisciplinary team of domain experts and visualization researchers further informs the design and evaluation of our approaches.

As illustrated in Fig. 1(d), our first approach was to enrich the numerical sensitivity measures with detailed visualization depicting the impact of different values of individual parameters on the model outputs. This enabled domain experts to observe the sensitive behavior of each parameter as estimated scalar functions instead of individual scalar values; to confirm, interpret, or raise questions about the numerical sensitivity values; and differentiate sensitive behaviors for parameters with similar numerical sensitivity values. We refer to this approach as the *algorithm-centric and visualization-assisted approach*.

As illustrated in Fig. 1(e), our second approach was to enable domain experts to observe the interactions between parameter sets and non-scalar model outputs (i.e., time series), as well as among different parameters and different time series. Given the often large number of parameters, and the even larger number of interactions, it is challenging to study these relations and reason about them using solely the table and plot as shown in Fig. 1(a) and (b). We therefore introduced clustering algorithms to reduce the cognitive load for data grouping, similar to Xiang and Swallow [61], to enable domain experts to identify and explore subsets of model outputs featuring similar behaviors in conjunction with the corresponding parameter sets. This enables domain experts to integrate their own domain knowledge in the reasoning process on the parameter sensitivity, e.g., about the semantic meaning of the visual patterns in a cluster of time series, or the scientific meaning of a specific range of a parameter. We refer to this approach as the *visualization-centric and algorithm-assisted approach*.

In addition to presenting the two visual analytics approaches, we report on our user-centered task analysis and problem-driven development and evaluation of both approaches. Our study showed that visual analytics can significantly improve the algorithm-only approach common in SA, and that domain experts welcomed the two approaches proposed in this work, which complement each other in SA.

In summary, our main contributions are as follows: (i) two complementary visual analytics approaches to support SA in epidemiological modeling as part of a large-scale collaboration between visualization and domain experts [12, 20] (both approaches have been integrated into the RAMPVIS server [29, 42] resulting from this collaboration, making them available to the modeling community at large); (ii) a user-centered task analysis and evaluation, and a mapping of the relations between the tasks and relative merits of these two approaches.

2 Related Work

Developing a simulation model is a highly iterative process, involving many techniques, such as ensemble simulation, uncertainty analysis, sensitivity analysis, dynamic steering, model inspection, model tweaking, and so on. The first three topics are often grouped under the parent topic of *uncertainty quantification* where uncertainty is propagated through all modeling stages [16]. In this context, ensemble analysis (prediction uncertainty) is concerned with studying the distribution of outputs while sensitivity analysis (attribute uncertainty) studies the relationship between input and output.

In this paper, we concentrate on visualizing the parameter space as a way to explore the ensemble data and sensitivity. Both the *algorithm-centric, visualization-assisted* and *visualization-centric, algorithm-assisted* approaches we explore in this paper integrate visualization with algorithmic approaches as proposed by Bertini and Lalanne [4].

2.1 Ensemble visualization

Ensemble visualization is an active research area concerned with providing a better understanding of ensemble data produced by computational simulation models [58]. Over the last decades, different research domains, such as engineering [35, 37], graphic design [8, 25], meteorology [5, 11], and epidemiology [15, 54], have developed and established theoretical and mathematical foundations based on their prior knowledge of ensemble data from models with various sets of model configurations. This ensemble data reflects unobserved reality at some unknown or extreme conditions. Simulations in engineering and graphic design applications benefit from ensemble data at the level of the design process. Trial and error experimentation can be reduced, lowering costs and increasing efficiency. While for other simulated data, e.g., adversarial weather events [17] and storm path ensembles [33], analyzing data improves timely risk mitigation planning when real data cannot be collected due to limited time. The epidemic modeling and decision support system from Afzal et al. [1] is closest to our work, where a visual analytics approach enables experts to evaluate a number of potential modeled responses to the pandemic. This work did not explore the relation between parameter ranges and outcomes systematically, which we emphasize. Our work falls in-between the above application categories, and aims to support the design and fine-tuning of models through visualizations of ensembles in relation to their parameter configurations.

Uncertainty analysis is one of the main tasks in ensemble visualization and a key design concern for our project, where we aim to communicate a range of model outcomes in compact representations. Hummel at al. [27] proposed a color mapped visualization to demonstrate variances of Lagrangian neighborhoods to capture and analyze flow instability. Contour boxplots [60], a generalization of boxplots, were introduced to characterize the uncertainty of feature sets in fluid dynamics applications. The EnConVis [60] tool was proposed to provide a unified framework for ensemble contour visualization. Specifically, ensemble data was clustered and its distribution modeled via Kernel Density Estimation before contour plots were used to visualize the ensemble uncertainty. The Noodles [49] tool leveraged spaghetti plots and bespoke glyphs to explore uncertainty in how water-vapor mixing ratio, perturbation potential temperature and perturbation pressure influenced the 1993 "Super-storm" phenomena. Our work extends this space by utilizing uncertainty visualization techniques in relation to clustered model outputs and associated parameter ranges.

Sophisticated algorithmic and visualization-based methods have been explored to tackle the challenge of simulated data embedded in a complex and high-dimensional space. Parallel Coordinate Plots (PCPs) are a popular option for visualizing high-dimensional ensemble data [3, 11, 31, 37]. For instance, Kumpf at al. [31] proposed a cluster-based brushing operation to visualize parameter distributions of ensemble members. After the brush is applied to all the ensemble data, multi-parameter violin plots are deployed to further analyze the patterns of clusters in each output parameter axis. Chen et al. [11] introduced a multi-view uncertainty visualization tool integrating PCP, a geo-location view and uncertainty histogram to explore ensemble data-sets. Clustering methods are also commonly deployed in many systems. Hao et al. [23] used agglomerative clustering to build a hierarchical tree for temporal ensemble data analysis. In work by Kumpf et al. [32], meteorological ensemble data was clustered by applying the k-means algorithm to the Principal Component Analysis (PCA) features and the variability of selected clusters visualized to identify representative data trends. Our work incorporates clustering, but does so on the model outputs to identify similar outcomes and introduces a novel visualization to reveal patterns within the parameter space.

2.2 Sensitivity Analysis

SA is *"the study of how uncertainty in the output of a model (numerical or otherwise) can be apportioned to different sources of uncertainty in the model input"* [44]. SA provides value for modelers by confirming and revealing how features of interest of the model output are affected

by variations or uncertainty in the inputs. Information on model sensitivities enable the modeler to make informed decisions as to what aspect of the model to develop further, and which model parameters to constrain with real-world data for more accurate forecasting.

SA methods can be broadly classified into local and global methods. Local methods measure the effect of small perturbations of parameter settings (i.e., gradients). Global SA measures compute an average effect across the range of a parameter. This effect can be measured in several ways including, Sobol Sensitivity indices [51], Fourier analysis [48], derivatives [52], among others [6, 38].

It is considered good practice [46] to use global measures of sensitivity rather than local ones, as local ones only show effects around a limited set of parameter settings. Furthermore, it is difficult to evaluate interaction effects using local analysis. In global SA, measures of sensitivity based on the variance of the model output are often used. Such "variance-based measures" are not conditional on the additivity or linearity of the model and are able to capture the effect of interactions between parameters [47]. As discussed later in Section 4, we make use of variance-based global measures in our approach.

Visualization of global sensitivity is challenging due to the dimensionality of the manifold that one wants to understand. As global sensitivity metrics produce a set of scalar sensitivity values for each parameter (and combinations thereof), domain experts often use bar graphs to visualize effects [19]. Visual methods for SA broadly fall into two camps: interactive methods and function plots. Interactive methods, such as the system by Matkovic et al. [36] or Berger et al. [3], enable the user to use brushing and filtering to explore sensitivities. These require discrete samples of the, often continuous, model and present *local* sensitivity. Extensive exploration is needed to build an understanding of the global sensitivity of the model and interaction effects are difficult to evaluate. Interacting with brushes and seeing how filtered results change, enables the user to build an intuition of the sensitivity of the model. For example, Guo et al. [21] use star glyphs around a set of sample points to show local sensitivities around each sample with respect to each parameter. Huang et al. [26] instead show how the local sensitivity around parameter values influences the coloration in choropleth maps. The PeTTSy [18] modeling tool is designed with compartmental epidemiological models in mind. It uses a SVD of the parameter space to show local sensitivities as heatmaps and line plots. Drag and Track [39] uses PCA and MDS to produce 2D views of the input and output parameter spaces. Users can get a sense for the local sensitivity by manipulating a cursor in one view (input or output) and seeing the effect in the other. Most of these approaches focus on local sensitivities and relies on the skills of the analysts to draw conclusions on the global sensitivities. Our work addresses these through measures and visualizations that operate at a global level.

Function plots visualize the multi-dimensional manifold directly, maintaining the perception of the model being continuous using the familiar notion of a function plot, often using one function plot per dimension. These methods present more of a *global* view of the parameter space. In addition, these function plots can show the "shape" of the effect, giving additional insight beyond scalar global sensitivity metrics. For example, Tuner [56] uses a set of 2D heatmaps to show sensitivity around a particular point in parameter space. The "sensobol" [41] R package can produce various plots of the outputs of variance-based SA and scatter plots showing the raw output data against a parameter. Contribution to the sample mean plots [6], show how the average output converges as the number of samples increases. Like the numerical methods, these illustrate average behavior of the manifold. SA of epidemiological data, including dimension reduction and temporal clustering, have successfully extracted important dynamics from multivariate data on COVID-19 dynamics (e.g., [54, 61]) to facilitate visualization and inference. The global sensitivity scatterplot [10] visually integrates local derivatives which are comparable to derivative-based global sensitivity measures. Many of these methods, such as principal components analysis, are geometrically linked to variance-based sensitivity measures.

In this paper we explore using a number of slices of the manifold [57], as they are intuitive to understand and can show more detail than the purely numerical or contribution to the sample mean plots. We were able to extend this method to handle time-varying outputs. Previous work on both visualization and numerical methods of SA only focuses on scalar outputs. However, epidemiological produce time series outputs which have non-scalar features, such as different shapes.

3 Sensitivity Analysis in Epidemiological Modeling

The SARS-CoV-2 pandemic has generated a prevalence of temporal trajectories of epidemic waves and progression indices that provide a challenging case study for the proposed approaches. With a significant increase in the number of mechanistic and compartmental models to simulate the pandemic under varying conditions, simulation outputs under these varying conditions and parameterizations have also grown considerably. Here we detail the epidemiological modeling frameworks used to demonstrate the proposed visualization approaches.

The principal epidemiological model used for the application of the proposed approaches is a stochastic forward simulation model of COVID-19 dynamics in Scotland [40]. The simulator is based on an age-structured compartmental SEIR model, with hospitalization and mortality. The model was specifically developed for forward simulation of the first wave of the epidemic in Scotland, focusing on the impact of interventions to harbor the spread of the disease in Scotland, particularly amongst care workers and medical professionals, and aiming to account for and/or estimate the impact of asymptomatic cases. It also has the capacity for calibration to observed data using Approximate Bayesian Computation (ABC-SMC). The model was used as a tutorial on Uncertainty Quantification, SA and calibration as part of the SCRC response [19], from which the domain knowledge was generated. The model was referred to as the EERA model [40].

A second model, the Simulation.jl model, is also used to test the visualization techniques in this work. It is a redeveloped version of the model for COVID-19 dynamics by Harris et al. [24]. This is an age-structured stochastic forward simulation model, developed to simulate the spread of COVID-19 in Scotland on a high-resolution spatial grid.

These models were selected as they represent two extremes of the epidemiological response in Scotland. The principal model is a high-aggregation model developed for studies of COVID-19 transmission within health board settings, and to answer questions on asymptomatic transmission and silent spread during the pre-detection period. The study of this model, combined with the Irish modeling approach, highlighted a modeling structure of the COVID-19 dynamic flow within a population which was then implemented within the Simulation.jl model. The latter model is a high-resolution spatial model which enables more complex studies of local transmission and environmental drivers.

SA is resource hungry in terms of time and computation, so its formal use in epidemiological emergencies has been limited by the speed at which decisions need to be made. Much previous work relies on forward simulation of models and informal visualization of outputs. In real-time this is challenging [34], Dunne et al. [19] discuss how to formally analyze models and make them more accessible. However, SA is critical to better understand complex relationships that might exist between parameters and affect the quality of inferences made using models [53]. This benefit outweighs the cost of carrying out SA.

4 User Tasks in Sensitivity Analysis

We began the project in June 2020 as part of an effort to leverage visualization and visual analytics to provide rapid assistance in modeling the pandemic (RAMP) [43]. Through biweekly meetings and discussions with domain experts for more than a year, described in detail in Section 7, we identified the following user tasks for SA.

Global or Overall Sensitivity:

a. To determine the sensitivity of each parameter.
b. To compare and order parameters according to their sensitivity.
c. To confirm if the algorithmic estimation of the sensitivity is reasonably meaningful or identify any questionable aspects of the algorithmic analysis (e.g., anomalies).
d. To acquire the characteristic patterns of sensitivity, e.g., linearity, consistency and sampling sparsity.

Sensitivity related to specific visual patterns in the model outputs:

e. To determine pattern-sensitivity of each parameter.
f. To compare pattern-specific sensitivity.
g. To identify which variable contributes to acceptable or unacceptable patterns.

Sensitivity related to different value ranges of each parameter:

h. To determine range-sensitivity of each parameter.
i. To compare range-specific sensitivity.
j. To identify which variable increases or reduces output variance, and does so in which specific value range.

Sensitivity related to multiple parameters:

k. To hypothesize the combined sensitivity of two or more variables.
l. To test a hypothesis about some combined sensitivity.
m. To analyze pattern- and range-specific combined sensitivity.

Furthermore, in sensitivity analysis literature [45], the main use cases for algorithmic sensitivity analysis are presented as four "settings":

- Factors Prioritization (FP): Identifying the parameter which, if determined, leads to the greatest reduction in the variance of the output. This involves ranking the parameters in order of importance for reducing the uncertainty in the output.
- Factors Fixing (FF): Identifying parameters with little effect on the output, which can be fixed or eliminated. Thus, reducing the complexity of the model.
- Variance Cutting (VC): What parameters should be better determined to decrease the variance beneath a certain threshold.
- Factors Mapping (FM): Addresses "what parameter is most responsible for producing outputs within a region of interest?".

4.1 Condensed List of Tasks for Visualization

While many epidemiological models (such as those used in this work) produce time series outputs, algorithmic sensitivity analysis studies scalar features of the output time series, e.g., the sum of values or the value at a specific time. Although algorithmic sensitivity analysis can only study scalar features, domain experts are also interested in non-scalar features, such as a time series' shape. For the methods developed in this paper we will focus on two features of the outputs of the EERA model, which domain experts deemed epidemiologically meaningful:

(A) **The peak number of deaths during the first 200 days**, providing an indication of the maximum strain a healthcare system would experience. A metric commonly used to approximate inference in simulation models (including Simulation.jl).

(B) **The overall shape of the time series of daily deaths during the first 200 days**. This measure enables modelers to identify outputs which match real-world data and outputs which deviate in interesting ways.

(A) is a scalar quantity while (B) is a qualitative feature of the output, of interest to the domain experts consulted. Visual encodings, as opposed to a purely algorithmic approach allow for the study of qualitative features of a model's outputs. The analysis and design of visualizations for alternative scalar and time series outputs would proceed identically to that of outputs (A) and (B).

A domain expert asked if we could introduce means of observing combined sensitivity, i.e., tasks (k,l). It was decided that allowing the user to supply a list of hypothesized interactions (functions of the model parameters) would be a suitable solution. This solution enables studying interactions involving an arbitrary number of parameters. Furthermore, showing the hypothesized interactions alongside the actual model parameters allows the user to visually determine the significance of the interaction using the same visual cues as for the other parameters. An alternative approach [41] is to encode the model output as a heat map where the coordinates encode values of 2 parameters, this does not easily extend to interactions involving more parameters.

Combining the list of tasks (a-m) and the settings listed in Section 4 with our output features of interest, we consulted with domain experts to settle on a condensed list of realistic sensitivity analysis tasks for an epidemiological model. The tasks are:

1. Rank the importance of the model parameters for the value/shape of the output.
2. Rank the importance of the model parameters in reducing the variance of the output.
3. Identify sensitivity patterns with regards to:
   - 3.1 Over what range parameters affect outputs.
   - 3.2 Parameters which affect the outputs when other parameters take certain values.
4. Identify whether hypothesized interactions based on prior knowledge correlate with the outputs as expected.

Where in the context of the epidemiological model, task 1 corresponds to the FP, FF, FM settings and tasks (a,b,c), task 2 corresponds to FP, FF, VC settings and tasks (a,b), tasks 3.1 and 3.2 corresponds to the FM setting in this context and to tasks (d,e,f,g,h,i,j), task 4 has no analogue in the "setting" but corresponds to tasks (k,l,m).

4.2 Task Driven Visualizations

Three visualizations were developed, each corresponding to a different approach to sensitivity analysis. The approaches and visualizations are:

Visualizing results of a sensitivity analysis algorithm.

(I) A bar chart showing the Sobol indices for scalar outputs.

Algorithmic-centric visualization assisted approach.

(II) A scatter-line plot of parameters values and scalar outputs, featuring 1D slice plots from a Gaussian process emulator (GPE) showing the expected effect of varying a single parameter.

Visualization-centric algorithm assisted approach.

(III) A cluster-based technique which computes clusters in the input or output space of the model, to reduce the cognitive load for data grouping when observing the data.

A modeler highlighted that sensitivity analysis and parameter tuning are often done in parallel. This involves searching for parameter values which result in outputs matching real-world behavior. The need to identify parameters for which the model outputs match real-world data motivates representing the model output using the same quantities used to quantify the real phenomena. Because of this, dimensionful and unnormalized model outputs were preferred, such as total deaths rather than deaths as a percentage of the total population. This also motivated not representing model outputs using algorithmically selected features, such as PCA components. In addition, according to the information theory justification of visualization [14], visualization adds value by allowing the observer to combine the shown data with their relevant prior knowledge. Representing the model outputs using quantities and visual encodings familiar to the modeler reduces the cognitive load required to compare model outputs with their prior knowledge of the phenomena being modeled.

It was decided to keep all axis labels consistent with the names used in the model. The choice was made to maintain continuity with the naming conventions used by the domain experts. We privileged the familiarity the domain experts have with the data, down to the meaning carried by the choice of specific names, over attempting to modify names and labels to perhaps increase transparency for a general audience but with the risk of potentially introducing mistakes.

4.3 Sobol Indices for Sensitivity Analysis

In this paper we use *Sobol indices*, a form of variance-based global sensitivity analysis, to quantify the influence of sets of parameters on the model output. See Saltelli [45] for the underlying mathematics. The intuition behind Sobol indices is that they quantify how much the variance in a model's output is expected to decrease when a parameter is fixed. The Sobol indices are normalized.

Fig. 2 shows a set of Sobol indices, including the first order (main effect) and interaction Sobol indices of the parameters, whose meanings are shown to the right, with respect to a scalar output of the EERA model. The interaction Sobol index quantifies the magnitude of interactions between one parameter and all the others. This visualization was developed to address task 1 for output A. While Sobol indices are scalars, and can be ordered algorithmically, the visualization provides a quick overview of the sensitivities present in a model using quantities familiar to SA practitioners.

This article has been accepted for publication in IEEE Transactions on Visualization and Computer Graphics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2022.3209464

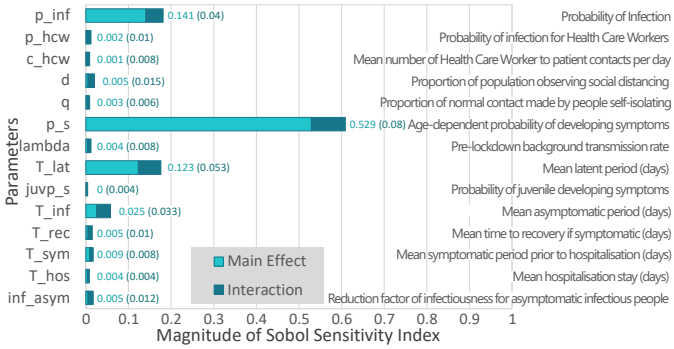

Fig. 2. A Sobol index plot showing the main effect and interaction effect of the parameters in the EERA model with respect to the maximum number of daily deaths during the first 200 days.

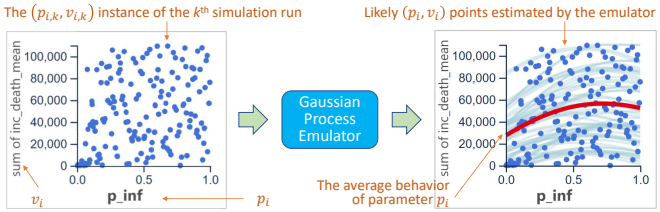

Fig. 3. The interpretation of a scatter plot for summarizing all simulation runs can be significantly improved by introducing lines to the plot, where each light blue line represents the trend of a run when it is extended by extra real or emulated runs where only one parameter varies.

It is not feasible to include the hypothesized interactions from task 4 in Sobol analysis. This is as Sobol analysis assumes that the model parameters can take on values independently from one another, but hypothesized interactions are functions of model parameters. Therefore, hypothesized interactions cannot be included in Sobol analysis.

5 Algorithm-Centric and Visualization-Assisted

While analytical algorithms, such as Sobol indices, can perform tasks (a) and (b) in Section 4, they do not support tasks (c) and (d) as they reduce information too quickly to aid these tasks. One solution to address this shortcoming is for users to visualize data that is not shown in the algorithmic results, such as the bar chart in Fig. 2 [13]. For example, when a user observes estimated sensitivity values of the parameter p_inf in Fig. 2 and notices the similarity with T_lat, the user naturally wishes to know if the global sensitivity of these two parameters really is similar. In this section, we present an approach to assist the interpretation of the algorithmic results through visualization.

Let $P = \{p_1, p_2, \ldots, p_n\}$ be a set of input parameters, $T = \{t_1, t_2, \ldots, t_m\}$ be a time series output, and S_{in} and S_{out} other input and output data. A model $\mathbf{M}$ is thus a function $(T, S_{\text{out}}) = \mathbf{M}(P, S_{\text{in}})$. The k-th simulation run of the model takes an instance parameter set $P = A_k$, and generates an instance output time series B_k, such that

$$(\{\beta_{1,k}, \beta_{2,k}, \ldots, \beta_{m,k}\}, S_{\text{out},k}) = \mathbf{M}(\{\alpha_{1,k}, \alpha_{2,k}, \ldots, \alpha_{n,k}\}, S_{\text{in},k})$$

The output of a simulation run can be characterized by one or a few statistical measures. Consider a simple measure v^{sum}, the sum of $\{t_1, t_2, \ldots, t_m\}$. The bi-variate relation (p_i, v_i^{sum}) informs us on the behavior of each parameter p_i $(i = 1, 2, \ldots, n)$ and thus its sensitivity. The visualization on the left of Fig. 3 shows the (p, v^{sum}) pairs for parameter p_inf obtained from 160 simulation runs. Although there does not seem to be any correlation, we cannot draw such a conclusion because different dots (p, v^{sum}) feature different parameter sets, hence an extensive amount of confounding effect.

Ideally, we could observe the behavior of p_i while fixing all other parameters p_j $(j = 1, 2, \ldots, n; j \neq i)$. However, this would be intractable to both computation and visualization. Since there is already a set of simulation runs with parameter sets A_k $(k = 1, 2, \ldots, l)$, the user can focus on individual runs. This translates to observations based on each instant parameter set $\{\alpha_{1,k}, \alpha_{2,k}, \ldots, \alpha_{n,k}\}$. The behavior of p_i under the condition of $\alpha_{j,k}$ $(j = 1, 2, \ldots, n; j \neq i)$ thus informs us of the sensitivity of p_i under that condition.

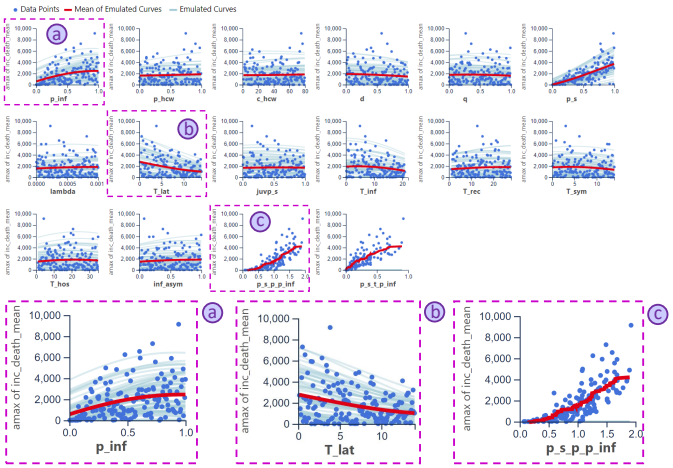

Fig. 4. Scatter-line plots for the EERA model showing 14 parameters and two combinations of parameters with respect to the maximum number of daily deaths. Zoomed-in views of p_inf (a) T_lat (b), and p_s + p_inf (c).

This suggests that for each data point (p_i, v_i) in the left scatter plot in Fig. 3, one may use a brute-force method to invoke more simulation runs to sample p_i under the fixed condition of $\alpha_{j,k}$ $(j = 1, 2, \ldots, n; j \neq i)$. However, the computational cost would be unattractive as one would have to compute these additional runs for all n parameters and all l existing runs (i.e., all dots in the left scatter plot in Fig. 3).

One can reduce the cost of additional simulation runs by emulating such runs using a Gaussian process emulator (GPE) [28], generating likely (p_i, v_i) points when p_i varies in its range (i.e., [0, 1] in this case). These likely points are shown as light blue curves in the combined scatter-line plot on the right of Fig. 3. Each curve in the scatter-line plot depicts the effect of just varying p_i while fixing the other parameters p_j $(j = 1, 2, \ldots, n; j \neq i)$ to constant values sampled uniformly.

A GPE is trained on the data of known simulation runs, i.e., $A_k, B_k)$ $(k = 1, 2, \ldots, l)$, and is used to predict the expected effect of varying the parameter values. For simulation runs that sample parameters using a uniform sampling scheme, such as Saltelli's extension of Sobol sequences [9] or Latin-Hypercube sampling, the points with adjacent values of p_i can be inferred effectively from different values of p_j indirectly by the GPE. As our testing shows that 94.53% of the predictions made by the GPE fall within the 95% confidence interval, it is unlikely that any light blue line in Fig. 3 would deviate far from the true line. These light blue lines thus enable scientists to observe trend patterns that the normal scatter plot cannot convey.

In addition, the mean of the curves predicted by the GPE is depicted by the red curve in the scatter-line plot, indicating the overall trend $v_i^{\text{sum}} = f_i(p_i)$. The curved pattern of the red line in Fig. 3 indicates that the sensitivity of the parameter changes within the range [0, 1]. For example, based on visual observation, one may consider that p_inf is more sensitive in the range [0, 0.5) than [0.5, 1]. One can also observe the changes in the correlation between p and v^{sum}, enriching the interpretation of the sensitivity values. Therefore tasks (h), (i), and (j) in Section 4 are also supported by this approach.

5.1 Example of Use: Examining Similar Sobol Indices

Fig. 4 shows a set of scatter-line plots for assisting the interpretation of the sensitivity values of the EERA model shown in Fig. 2. The characteristic measure v in this figure is the maximum value of the output time series, i.e., $v^{\max} = \max(\beta_1, \beta_2, \ldots, \beta_m)$. From Fig. 2, we can answer the query about the two parameters, p_inf and T_lat. Although they have similar sensitivity values, the curves in the two corresponding scatter-line plots are noticeably different. The mean curve of (p, v^{sum}) for $p =$p_inf moves upwards when increasing p_inf, while that for $p =$T_lat moves downwards when increasing T_lat.

The original scatter plot can also display interactions between two or more parameters as shown in Fig. 4(c). Such multi-parameter plots do not have lines for emulated data, since the emulation would result in a 3D surface. The red line indicates the mean of the simulation runs.

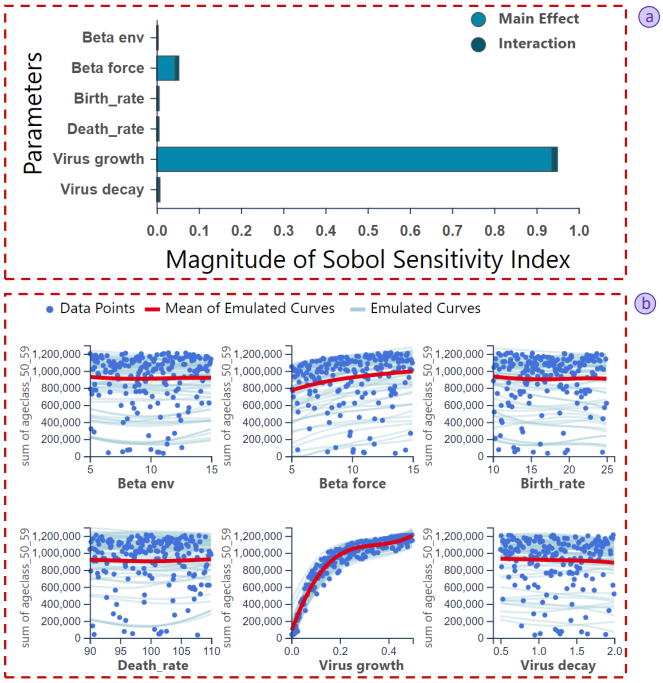

Fig. 5. Sobol index plot (a) and scatter-line plot (b) for the Simulation.jl model with respect to total number of deaths. Beta_env is the environmental transmission, Beta_force the force of infection, other parameters are explained by their names

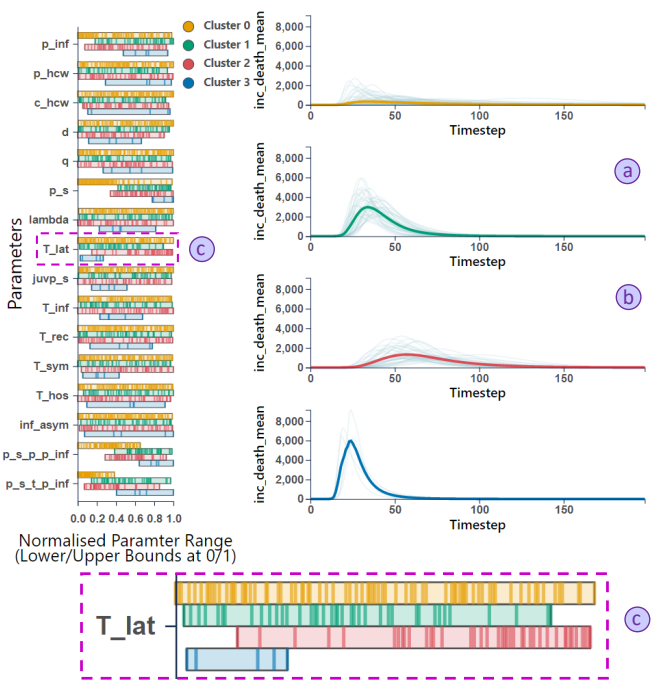

Fig. 6. Cluster-based sensitivity analysis plot for the epidemiological model. Two output clusters with an early sharp outbreak (a) and a later outbreak with a lower peak (b). The latent period, T_lat (c) is on average greater for the model runs associated with (b) than for (a), while the other model parameters have similar distributions for the two clusters.

5.2 Example of use: Analyzing the Simulation.jl Model

This use case investigates the effect of the parameters on total deaths in the ages 50-59 over a 2-month period. The analysis starts by generating a Sobol index plot, seen in Fig. 5(a). If one has little prior knowledge of a model's sensitivities, the Sobol index plot can provide a quick overview of the relative influence of the different parameters and an estimate of the importance of interactions. From Fig. 5(a) the parameter "Virus growth" appears to be most influential, with "Beta force" having some effect, and the rest having little effect. Furthermore, Fig. 5(a) suggests that interactions between parameters are of limited importance.

The scatter-line plots for the data set, seen in Fig. 5(b), should now be consulted. Fig. 5(b) shows that increasing the value of "Virus growth" only has a clear effect in increasing the number of deaths when it is in the lower part of its range, which the Sobol indices cannot reveal. Furthermore, the scatter-line plots show that "Beta force" has a slight positive correlation with the total deaths over its entire range.

6 Visualization-centric and Algorithm-assisted

In many applications, estimating global or overall sensitivity is usually not sufficient for understanding the behaviors of models and their parameters. Modeling scientists often visualize simulation results, select one or a few visual patterns that they would like to have or avoid in the results, and attempt to identify parameters that are sensitive to these visual patterns. In such situations, the sensitivity related to specific visual patterns, which are tasks (e), (f), and (g) in Section 4, may take the priority over the global sensitivity. As discussed briefly in Section 1, it would be cognitively challenging to directly observe the data of ensemble runs, which contain large amounts of information, to reason about the sensitivity of the parameters effectively.

For example, hypothetically, one could identify all time series in Fig. 1(b) that exhibit a certain visual pattern, and then observe if these time series correspond to some data patterns in the parameter table in Fig. 1(a). The data patterns would indicate how the parameters are sensitive to the presence or absence of the visual pattern concerned. According to an empirical study on five visualization tasks by Borgo et al. [7], humans' performance involving two concatenate visualization tasks can be significantly worse than that for each task individually. Although the viewing of the table in (a) or plot in (b) independently may be cognitively demanding but feasible, concatenating the two (i.e., viewing (a) with the observation of (b) in mind) would be in the challenging spectrum indicated in Borgo et al.'s research [7]. For such a costly visualization-centric approach, statistics and algorithms can provide rapid abstraction to reduce the cognitive cost [13].

In this work, we use a clustering algorithm to assist the grouping of time series in (b), cluster-based visualization as shown in Fig. 1(e) can provide the task of viewing the table in (a) with helpful abstraction and external memorization of the visual patterns. In addition, using the pixel-based visualization as shown in Fig. 6(c) can help to further reduce the cognitive load associated with viewing the parameter table in Fig. 1(a) by enabling users to observe the data patterns of each cluster within the range of each parameter.

Recall the notation in Section 5. In each simulation run, the model $\mathbf{M}$ takes an instance input parameter set (i.e., $P = A_k$) and generates an instance time series (i.e., $T = B_k$) as a prediction. Here we focus on the sensitivity of P in relation to the visual patterns in T, though there may be other forms of inputs ($S_{\text{in},k}$) and outputs ($S_{\text{out},k}$).

We can divide the l time series $B_1, B_2, \ldots, B_l$ into g clusters $C_1, C_2, \ldots, C_g$ based on a set of similarity criteria. As there are usually only a few clusters, i.e., $g \ll l$, cluster-based observation and reasoning allows users to "outsource" the mental effort of grouping and memorization to the algorithm and visualization, making more cognitive resources available to other related tasks, e.g., considering what visual patterns are important but not predicated by all simulation runs. As shown in Fig. 6, each cluster is represented by a mean curve as its "signature curve", allowing users to reason more easily as to which clusters feature a desired visual pattern and which do not. When the users move to the next task of observing the sensitivity of the parameters, instead of considering many curves that have similar visual patterns identified and stored in one's mind, the users can focus on a few signature curves in the external memory (e,g., four in Fig. 6) which, while limited to storing only a few objects at a time, is highly efficient [59].

The cluster-based sensitivity analysis plots consist of two parts, one showing the model parameters and another showing the output time series. To the left, a grouped chart shows the ranges and values each parameter takes for the model runs associated with that cluster. To the right, line plots of an output quantity against time is shown. The line plots contain translucent lines for all ensemble members in a cluster and

This article has been accepted for publication in IEEE Transactions on Visualization and Computer Graphics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2022.3209464

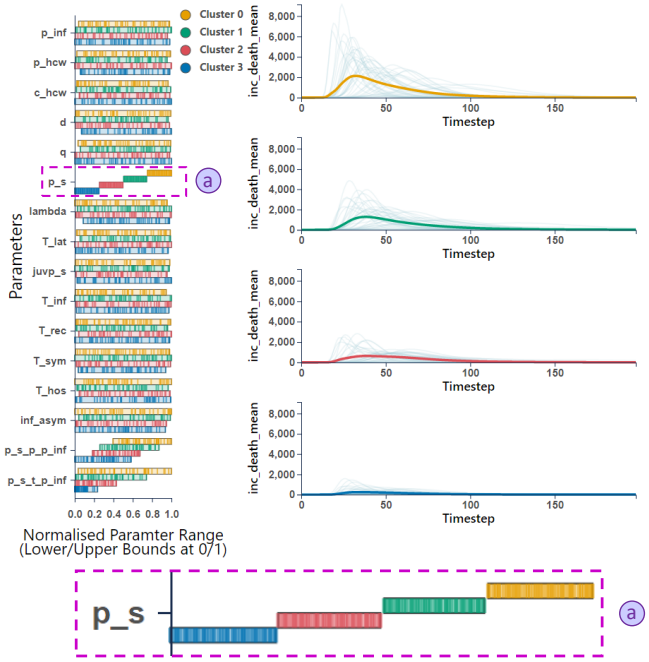

Fig. 7. Clustering based on the value of p_s (a) in the EERA model.

a thicker colored line representing their mean. To aid in matching the input and output spaces associated with a cluster, the color of the mean line in the line plot matches the colors used for the different clusters in parameter space. The vertical ordering of the bars in the parameter space plot is also consistent with the vertical ordering of the line plots.

All the lines in the ensemble are shown, rather than a confidence interval, to allow the modeler to identify outliers in a cluster and anomalous model behavior. A modeler can utilize the view of all lines in a cluster to identify anomalies and judge whether they might stem from the model behaving unexpectedly or from poor clustering.

Clustering can be based either on model input parameters, or output curves. To investigate the effect of a parameter or interaction taking on low, medium, or high values, the data can be clustered, so all model runs where a parameter or interaction takes values in the same $1/k$ part of its range are assigned to the same cluster. Fig. 7 shows clustering with respect to the value of p_s.

Clusters based on the output curves can be computed using time series clustering algorithms [55] such as k-means based on Euclidean distance or dynamic time warping. Clustering output time series reveal whether outputs with certain shapes correspond to certain sets of parameter values. One can uncover interactions between parameters by comparing their values across clusters, see Section 6.2. In addition, output clustering allows domain experts to match model behavior with real-world time series using their experience, Section 6.1 discusses this. Output clustering can also provide a sanity check, e.g. to confirm that there is no or only a minor outbreak when the probability of infection is low, and deviations from expected behavior can motivate further study.

Determining the appropriate number of clusters, k, for a data set is a nontrivial task to perform algorithmically. Because of this, the modeler supplies a list of k values to be tried; utilizing their prior knowledge regarding the expressivity of the model, derived from knowing its inner workings, to narrow down the search space for the optimal k. The cluster-based analysis is then performed, producing plots, for all values of k supplied. When presented with the plots for the different k values the modeler can identify and focus on the plots which contain clusters which are epidemiologically meaningful or show interesting patterns in parameter space. Using visualization, the algorithmically complex task of finding an optimal k is replaced with the simpler task of identifying meaningful and interesting clusters in a handful of plots. Enabling a modeler to quickly determine if the plot for a k value is of interest is therefore a key requirement of this visual encoding.

Colors were selected to be distinct yet color blind friendly for most common types of color blindness, additional considerations were that the order should be such that adjacent colors differ substantially so that the different lines in the parameter chart to the left in the cluster plot are distinct when next to each other as seen in Fig. 7(a).

On the left-hand side, bars show the range each parameter takes in the different clusters with vertical one-pixel wide lines showing the individual parameter values present in the cluster. Vertical lines are used since they allow a user to gauge the density of the distribution, to identify outliers or a bimodal distribution of parameters in a cluster. The layout was compared against alternative options such as box-plots and violin-plots, and discussed with domain experts - who appreciated the advantages brought by less compression of the data and more details on the overall distribution, in a small form factor. The individual ticks within a bar are of a similar color to the bar itself to not break up the shape of the bar. In output clustering, parameters with similar distributions for many clusters, are not important for the shape. Clearly presenting the shape of the bar helps a user in quickly identifying parameters with little impact. Having discarded unimportant parameters, the user can use the colored vertical lines to study parameter distributions of interest further, as shown in Fig. 6(c).

6.1 Example of Use: Addressing Questions Relevant to the Initial COVID-19 Wave in Scotland

Domain experts highlighted that different questions have been relevant at different times and places of the COVID-19 pandemic. During the initial wave, questions of interest in Scotland included: the prevalence of asymptomatic spread and the amount of community transmission prior to the first lockdown on March 23rd, 2020. These questions can be addressed by constraining the values of "pre-lockdown background transmission rate" and "mean asymptomatic period" in the EERA model, by comparing outputs and real-world data.

A domain expert who used the EERA model to study this very problem at the time, proposed utilizing the cluster-based plot with clustering based on the output time series to study this. The domain expert used his prior knowledge of the shape of real-world time series of deaths in Scotland to identify two clusters of output time series whose shapes were in rough agreement with the shape of the real-world time series, these were (a) and (b) in Fig. 6. By identifying the parameter values associated with these clusters in the cluster-based analysis plot, estimates for the rate of spread before the initial lockdown in Scotland and the mean asymptomatic period could be obtained. In this way, domain experts can use the visualization-centric algorithm assisted approach and their knowledge of the real-world time series to quickly estimate the values of real-world quantities relevant to epidemiology.

6.2 Example of Use: Obtaining an Overview of the Sensitivities of the EERA COVID-19 Model.

Modelers with prior experience of the EERA model suggested that the value of $p_s + p_{inf}$ and $p_s \times p_{inf}$ (scalar multiplication) may be influential in determining the peak number of deaths. These were therefore included as hypothesized interactions. Fig. 2 shows the Sobol indices based on the value of peak deaths, suggesting that four parameters (p_{inf}, p_s, T_{lat}, T_{inf}) are most influential. The Sobol index plot agreed with the domain experts experience-based hypotheses.

In the Fig. 4 scatter-line plots, the model parameters whose red lines have the most pronounced gradient match the four most influential parameters from Fig. 2. Fig. 4(a) reveal that p_{inf} seems to determine the peak number of deaths only when it is towards the bottom of its range, since there the vertical spread of the blue dots corresponding the spread in peak deaths is much reduced. The values of p_s and p_{inf} correlate positively with the peak deaths while T_{inf} and T_{lat} (b) correlate negatively. Both $p_s + p_{inf}$ and $p_s \times p_{inf}$ correlate strongly with the peak number of deaths. However, the value of $p_s + p_{inf}$ (c) provides a tighter and roughly linear lower and upper bounds of the peak deaths.

One can find parameters which are important conditional on the values of the other parameters using output clustering. This is shown in Fig. 6 where most parameters associated with cluster (a) and (b) have similar distributions, including p_s taking on large values. However, the distribution of T_{lat} (c) is different in the two clusters, the range of

This article has been accepted for publication in IEEE Transactions on Visualization and Computer Graphics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2022.3209464

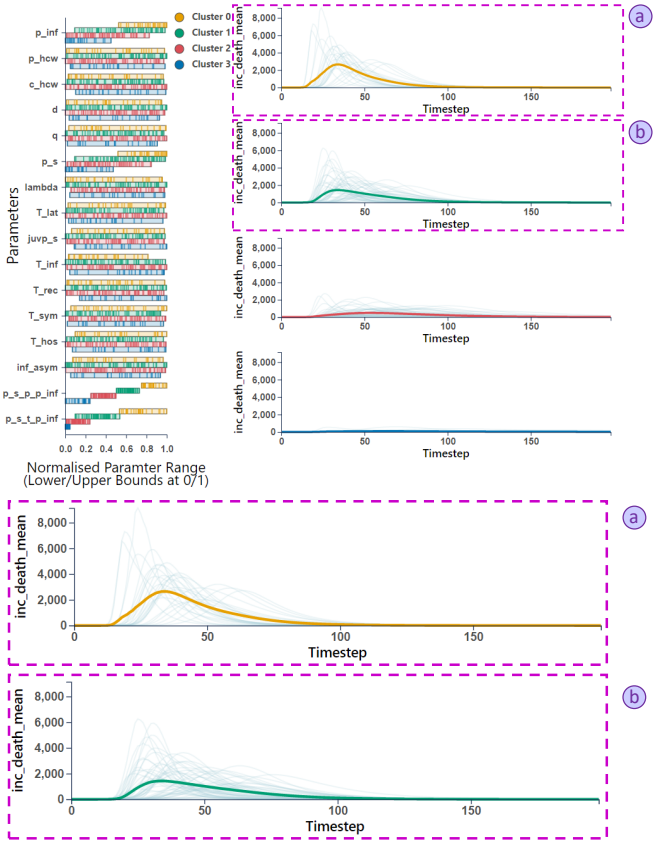

Fig. 8. Clustering for the principal epidemiological model, showing output time series corresponding to values of the hypothesized interaction $p_s + p_{inf}$ in the largest 25% (a) and second largest (b) 25% of its range.

the distributions overlap but their centers are in the lower and upper part of T_{lat}'s range respectively. This shows that in this model T_{lat}, the latent period, has a significant effect in determining the peak number of deaths if p_s is large enough for an outbreak to occur, otherwise T_{lat} has limited effect, as seen from its distribution in the yellow cluster, corresponding to a limited outbreak. Since the scatter-line plot showed that $p_s + p_{inf}$ correlated strongly with the peak number of deaths it was decided to cluster the output curves based on the value of $p_s + p_{inf}$, shown in Fig. 8. However, in Fig. 8 some lines in cluster (b), have peaks above lines in cluster (a), showing that the value of $p_s + p_{inf}$ alone does not predict the peak number of deaths. Although, on average the lines in the different clusters follow the expected trend.

This use case reflects the order domain experts preferred to use the plots in this work to get an overview of the sensitivities of a model. Starting with Sobol indices, then the scatter-line plot and finally cluster-based sensitivity analysis plots allow the user to carry forward insights from less involved visual designs to more involved visual designs.

7 User Evaluation

The project dates back to June 2020 and it was initiated as part of an effort to support domain experts in what could be described as *Guerrilla Analytics* [42]. The work can be classified as *problem driven* and the design process followed a similar structure to the design study methodology [50]. The design process pipeline traversed three main phases: an exploratory phase where we analyzed domain and problem space; a formative phase where designs were proposed and validated; and finally a summative phase which gathered reflections to confirm and improve our solutions. Biweekly meeting with members of the RAMPVIS community were held throughout the process to discuss our proposals and keep up to date with the domain experts' current work, gather epidemiologists' feedback and new requirements, and explore possible new directions. This was fundamental to identify new opportunities and validate the flexibility, relevance and impact of our work. We report details of each individual phase. The domain expert team included 6 Senior Research Scientists, 1 Research Fellow, 2 final year PhD students. The team was composed of: epidemiologists, epidemiology modelers and mathematical epidemiologists.

7.1 Exploratory Phase and Formative Evaluation

Our initial approach as data scientists and visualization experts focused on the exploration of the model input and output space and linkages across the two. The modelers looked favorably at our attempts and brought to our attention that a core component of their work concerns analyzing the input parameter space. The modelers identified as core the analysis of parameters' sensitivity as a means to understand parameters interactions and effects on the final output. We soon noticed how modelers used the words "sensitivity" and "uncertainty" interchangeably, as parameter sensitivity is tightly correlated to uncertainty in the model output. This interaction contributed to a steep change in our approach and focus of our work, and highlighted a difference in terminology when referring to "uncertainty". Our work shifted from the more traditional development of a visual analytics platform to the investigation of metrics and methodologies used by modelers in their analytical and exploratory process (as detailed in Section 4). This phase allowed gathering a set of domain questions of interest to the modelers:

- On model stochasticity: D1: How do the model outputs vary for a fixed set of parameter configurations?
- On model uncertainty in relation to inputs: D2: How does the variation in the model outputs relate to the parameters of the model? Which parameters are most important or least important? D3: How does the relation between the input parameters and model output variations vary for different categories, e.g., for different age groups? D4: What are second degree and higher degree interactions in relation to the uncertainty? D5: Can we find out more about the nature of the sensitivity – if you have lower values within the parameters, what would be the outcome and/or the characteristics of the uncertainty?
- On model outcomes and real-world observations: D6: History matching – how does the model outputs match reality? D7: What inputs configuration fit the model to the observed data? D8: Which outcomes lead to certain inputs?

All questions were being addressed by the modelers through the use of emulators followed by SA on the respective parameter configurations, D8 required instead what the modelers referred to as reverse SA.

In our formative evaluation phase, questions 1-8 were unpacked into the high level tasks detailed in Section 4: D1(h-j), D2(a-b), D3(e-g), D4(k-m), D5(h-j,c-d), D6(e-g), D7(c-d), D8(e-g). We looked at methods that would support analysis of high dimensional data such as parallel coordinate plots, pixel-based visualization and dimensionality reduction techniques to identify patterns and clusters. The modelers were interested in the possibility of moving from a holistic view, as the one provided by a pixel-based visualization, to a detailed view of each parameter space and behavior individually. They were also interested in exploring beyond the statistical properties of each parameter and look at interaction between them.

Feedback from this phase guided the development of the two frameworks involving cluster-based visualization, small multiples and scatter plots among others.

7.2 Summative Evaluation

In addition to the formative development, we evaluated our designs using a series of case studies (see Sections 5.2, 6.1, 6.2) and individual domain expert interviews. With the exception of one of the researchers, who is also a co-author, the participants did not contribute to the design and development of the tools, except for the requirements gathering in the exploratory phase. After an introduction to each design, participants were asked to explore each tool independently and articulate their thought process and observations. Tasks presented in Section 4.1 were used to inform open-ended questions according to the think-aloud protocol procedure (interview script provided in supplementary material). The interview was followed by a short debriefing where we asked participants if they had any suggestions on future developments or extra

features that could be added and details on how such improvement(s) could help their work. Each session took between 40 to 60 minutes.

The interview explored how helpful each plot was in **ranking** and **identifying** the: (i) importance of the parameters in reducing variance and/or determining values; (ii) ranges of significant parameters and/or dependencies of parameters; (iii) testing and validating parameters interactions; (iv) and for the cluster-based sensitivity analysis plot also how the aforementioned affected shapes of the time series. Participants were extremely engaged and discussion touched on all core points.

7.2.1 Summary of Findings

The feedback we received was overwhelmingly positive. The domain experts had diverse feedback for each plot.

Algorithmic-Centric and Visualization-Assisted. The Sobol index plot was the most familiar visual encoding for the domain experts, they highlighted how our proposed layout was a significant improvement over state of the art. Experts noted how well the magnitude of interaction could be seen and how it supported comparison across parameters. Experts added how the visualization helped in the analysis of parameters which had already been identified as potentially influential.

Experts were particularly positive about the scatter-line plot. They unanimously reported how well the visualization provided a summary of each parameter space and overall behavior. We were surprised by the fact that, without any prompt from us, they looked at the plot jointly with the Sobol index plot. The senior experts and one researcher noted how the scatter-line plot allowed the detailed exploration of parameters which appeared of interest in the Sobol index plot. One of the experts reported how some of the parameters had similar statistical values in the Sobol index plot but differed significantly in the scatter-line plot *"Just looking at the magnitudes you would see a very high p_{inf} and p_s but on their own you would not be able to tell that actually an intermediate value of T_{rec} was really bad for you."*. One of the senior experts commented *"With the second visualization [scatter-line plot], you can see which parameter you really need to watch."*. Another expert admitted how he was very interested in parameter calibration and thought this plot to be a great tool to compare a combination of parameters, and how he could use such results to validate which pair had behaviors closer to the outcome of other simulations he was running. The same expert reported how he would have been interested in creating plots depicting pairwise combinations of parameters similar to the $p_{\mathrm{f}} + p_{\mathrm{inf}}$ case in Fig. 4(c).

Visualization-Centric and Algorithmic-Assisted. Feedback on the Cluster-based plot was very encouraging. Experts reported how the visualization enabled temporal SA, one of the experts reported that the plot together with parameter interaction expressed through the scatter-line plots can *"help reduce the "uncertain" space"*. One of the experts found the aggregate curves very interesting however his feedback was biased by his distrust in clustering algorithms *"The aggregate curves, that is very interesting [...] the representation actually looks really useful. If I could find a clustering method I really trust, then this is a great way to see the spaces where things are happening: high infectiousness, etc. The peak here is telling a useful story. If I wanted to get into a space where we do not have a high peak then I can look at what parameter space I am in [in the clustering]."* Another expert provided a similar comment highlighting how *"From the curves one can quickly note the flat curves, that means no outbreak, and the sharpness of the curve is interesting"*. He also added how *"... see the clusters there? They tell us to stay away from this parameter space because is yielding unrealistic results [pointing to the corresponding curve]"*. Further suggestions from participants included to overlay real-life curves if available, and to be able to quickly see the scalar features of different curves, such as the peak height. One of the experts wondered if there could be a way to integrate spatial data to exploit the model geographical information. Experts who were more positive to the use of clustering algorithms found the clusters enabling to perform a very quick comparison of the model output with real-world data.

7.2.2 Discussion

The evaluation confirmed our hypothesis that the two frameworks supported the two ends of the analytical process. Beyond the positive reception we also gathered important insights from the domain experts. These include the role of a co-creation process which saw both communities involved in trying to tease out the real challenges and potentials in both domains. The framework did not aim to provide answers for the modelers but rather to support the process of teasing out important dimensions of the data by allowing the user to reduce the search space themselves. *"I think the point is that it tells you straight away what you need to know there and then you can move on because you need to focus on those four parameters that have, you know, significant impact and maybe interaction."* We also noted how "freeing" some of the cognitive load made experts to ask for more features. These represent interesting future directions as well as new challenges.

8 Conclusions and Future work

In this paper, we have presented two visual analytics approaches for supporting sensitivity analysis in epidemiological modeling workflows. Our experience of task analysis indicates the importance of user-centered requirements gathering in the exploratory phase of the work as well as understanding the existing solutions including the algorithm-only approach commonly used by the domain experts and the existing visualization solutions for other applications.

Both our proposed approaches focus on tasks that domain experts wish to perform in SA, which are not adequately supported by numerical outputs of an SA algorithm and the statistical graphics for the analytical outputs. The first approach, "algorithm-centric, visualization-assisted", improves the algorithm-only solution by letting users observe significantly more detailed data that likely contains extra useful information about parameter sensitivity, enriching the typical statistical graphics of the analytical outputs. The second approach, "visualization-centric, algorithm-assisted", complements the algorithm-centric methods by providing users with visualizations that users can interpret easily and relate quickly to other information (e.g., important patterns in the real-world data). Our user-centered evaluation confirms the relative merits of these two methods over the algorithm-only approach.

In general, the uses of both approaches are not limited to epidemiological modeling. The first approach can work in any situation where model outputs are scalar values or can be summarized by some statistical measures. The second approach was designed for complex model outputs, for which different clustering algorithms may be required for different data types. Our current implementation is applicable to any model with time series outputs.

The two approaches are currently implemented as automated agents and visualization plots in a visualization infrastructure [42]. We plan to bring these two approaches together into an interactive system, where experienced users can have more control of the secondary algorithms such as the GPE and the clustering algorithm. Likely new visualization techniques will emerge from the R&D of such a sensitivity analysis system. This will allow formal assessment of the relative merits of our methods and traditional SA pipelines used by domain experts. Building on some of the feedback from the evaluation, we will also explore how one can foster greater trust in the kinds of approaches we propose here, such as exploring different clustering methods, and providing interactive capabilities to relate different data aspects more effectively.

Acknowledgments

The authors thank Michael Dunne, Hossein Mohammadi and Richard Reeve for sharing their valuable insights during interviews. Authors would like to thank UKRI/EPSRC "RAMP VIS: Making Visual Analytics an Integral Part of the Technological Infrastructure for Combating COVID-19" (EP/V054236/1), Scottish Government Rural and Environment Science and Analytical Services Division, Centre of Expertise on Animal Disease Outbreaks (EPIC), French National Research Agency and Boehringer Ingelheim Animal Health France for support through the IDEXLYON project (ANR-16-IDEX-0005), the Industrial Chair in Veterinary Public Health, as part of Lyon VPH Hub.

REFERENCES

[1] S. Afzal, R. Maciejewski, and D. S. Ebert. Visual analytics decision support environment for epidemic modeling and response evaluation. In *2011 IEEE conference on visual analytics science and technology (VAST)*, pp. 191–200. IEEE, 2011.

[2] R. Aguas, L. White, N. Hupert, R. Shretta, W. Pan-Ngum, O. Celhay, A. Moldokmatova, F. Arifi, A. Mirzazadeh, H. Sharifi, K. Adib, M. N. Sahak, C. Franco, and R. Coutinho. Modelling the covid-19 pandemic in context: an international participatory approach. *BMJ global health*, 5(12), 2020.

[3] W. Berger, H. Piringer, P. Filzmoser, and E. Gröller. Uncertainty-aware exploration of continuous parameter spaces using multivariate prediction. In *Computer Graphics Forum*, vol. 30, pp. 911–920. Wiley Online Library, 2011.

[4] E. Bertini and D. Lalanne. Investigating and reflecting on the integration of automatic data analysis and visualization in knowledge discovery. *SIGKDD Explor. Newsl.*, 11(2):9–18, may 2010. doi: 10.1145/1809400.1809404

[5] A. Biswas, G. Lin, X. Liu, and H.-W. Shen. Visualization of time-varying weather ensembles across multiple resolutions. *IEEE transactions on visualization and computer graphics*, 23(1):841–850, 2016.

[6] R. Bolado-Lavin, W. Castaings, and S. Tarantola. Contribution to the sample mean plot for graphical and numerical sensitivity analysis. *Reliability Engineering & System Safety*, 94(6):1041–1049, June 2009. doi: 10.1016/j.ress.2008.11.012

[7] R. Borgo, K. Proctor, M. Chen, H. Jaenicke, T. Murray, and I. M. Thornton. Evaluating the impact of task demands and block resolution on the effectiveness of pixel-based visualization. *IEEE Transactions on Visualization and Computer Graphics*, 16(6):963–972, 2010.

[8] S. Bruckner and T. Möller. Result-driven exploration of simulation parameter spaces for visual effects design. *IEEE Transactions on Visualization and Computer Graphics*, 16(6):1468–1476, 2010.

[9] F. Campolongo, A. Saltelli, and J. Cariboni. From screening to quantitative sensitivity analysis. a unified approach. *Computer physics communications*, 182(4):978–988, 2011.

[10] Y.-H. Chan, C. D. Correa, and K.-L. Ma. The generalized sensitivity scatterplot. *IEEE transactions on visualization and computer graphics*, 19(10):1768–1781, 2013.

[11] H. Chen, S. Zhang, W. Chen, H. Mei, J. Zhang, A. Mercer, R. Liang, and H. Qu. Uncertainty-aware multidimensional ensemble data visualization and exploration. *IEEE Transactions on Visualization and Computer Graphics*, 21(9):1072–1086, 2015.

[12] M. Chen, A. Abdul-Rahman, D. Archambault, J. Dykes, A. Slingsby, P. D. Ritsos, T. Torsney-Weir, C. Turkay, B. Bach, A. Brett, H. Fang, R. Jianu, S. Khan, R. S. Laramee, L. Matthews, P. H. Nguyen, R. Reeve, J. C. Roberts, F. Vidal, Q. Wang, J. Wood, and K. Xu. RAMPVIS: Answering the challenges of building visualisation capabilities for large-scale emergency responses. *Epidemics*, 39:100569, 2022.

[13] M. Chen and D. S. Ebert. An ontological framework for supporting the design and evaluation of visual analytics systems. *Computer Graphics Forum*, 38(3):131–144, 2019.

[14] M. Chen and H. Jänicke. An information-theoretic framework for visualization. *IEEE transactions on visualization and computer graphics*, 16(6):1206–1215, 2010.

[15] T. J. Cocucci, M. Pulido, J. P. Aparicio, J. Ruíz, M. I. Simoy, and S. Rosa. Inference in epidemiological agent-based models using ensemble-based data assimilation. *Plos One*, 17(3):e0264892, 2022.

[16] Committee on Mathematical Foundations of Verification, Validation, and Uncertainty Quantification; Board on Mathematical Sciences and Their Applications, Division on Engineering and Physical Sciences, National Research Council. *Assessing the Reliability of Complex Models: Mathematical and Statistical Foundations of Verification, Validation, and Uncertainty Quantification*. The National Academies Press, 2012.

[17] C. V. F. de Souza, P. da Cunha Luz Barcellos, L. Crissaff, M. Cataldi, F. Miranda, and M. Lage. Visualizing simulation ensembles of extreme weather events. *Computers 'I&' Graphics*, 2022. doi: 10.1016/j.cag.2022.01.007

[18] M. Domijan, P. E. Brown, B. V. Shulgin, and D. A. Rand. Pettsy: a computational tool for perturbation analysis of complex systems biology models. *BMC bioinformatics*, 17(1):1–19, 2016.

[19] M. Dunne, H. Mohammadi, P. Challenor, R. Borgo, T. Porphyre, I. Vernon, E. E. Firat, C. Turkay, T. Torsney-Weir, M. Goldstein, R. Reeve, H. Fang, and B. Swallow. Complex model calibration through emulation, a worked example for a stochastic epidemic model. *Epidemics*, p. 100574, May 2022. doi: 10.1016/j.epidem.2022.100574

[20] J. Dykes, A. Abdul-Rahman, D. Archambault, B. Bach, R. Borgo, M. Chen, J. Enright, H. Fang, E. E. Firat, E. Freeman, T. Gönen, C. Harris, R. Jianu, N. W. John, S. Khan, A. Lahiff, R. S. Laramee, L. Matthews, S. Mohr, P. H. Nguyen, A. A. M. Rahat, R. Reeve, P. D. Ritsos, J. C. Roberts, A. Slingsby, B. Swallow, T. Torsney-Weir, C. Turkay, R. Turner, F. P. Vidal, Q. Wang, J. Wood, and K. Xu. Visualization for epidemiological modelling: Challenges, solutions, reflections & recommendations. *Philosophical Transactions of the Royal Society A*, to appear.

[21] Z. Guo, M. O. Ward, E. A. Rundensteiner, and C. Ruiz. Pointwise local pattern exploration for sensitivity analysis. In *2011 IEEE conference on visual analytics science and technology (VAST)*, pp. 131–140. IEEE, 2011.

[22] L. Hadley, P. Challenor, C. Dent, V. Isham, D. Mollison, D. A. Robertson, B. Swallow, and C. R. Webb. Challenges on the interaction of models and policy for pandemic control. *Epidemics*, 37:100499–100499, 2021.

[23] L. Hao, C. G. Healey, and S. A. Bass. Effective visualization of temporal ensembles. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):787–796, 2015.

[24] C. L. Harris, C. A. Cobbold, N. Brummitt, and R. Reeve. Dynamic virtual ecosystems as a tool for detecting large-scale responses of biodiversity to environmental and land-use change. arXiv 1911.12257, 2019.

[25] W. He, J. Wang, H. Guo, K.-C. Wang, H.-W. Shen, M. Raj, Y. S. Nashed, and T. Peterka. Insitunet: Deep image synthesis for parameter space exploration of ensemble simulations. *IEEE transactions on visualization and computer graphics*, 26(1):23–33, 2019.

[26] Z. Huang, Y. Lu, E. A. Mack, W. Chen, and R. Maciejewski. Exploring the sensitivity of choropleths under attribute uncertainty. *IEEE Transactions on Visualization and Computer Graphics*, 26(8):2576–2590, 2019.

[27] M. Hummel, H. Obermaier, C. Garth, and K. I. Joy. Comparative visual analysis of lagrangian transport in cfd ensembles. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2743–2752, 2013.

[28] M. Kennedy and A. O'Hagan. Bayesian calibration of computer models (with discussion). *Journal of the Royal Statistical Society*, B63:425–464, 2001.

[29] S. Khan, P. Nguyen, A. Abdul-Rahman, E. Freeman, C. Turkay, and M. Chen. Rapid development of a data visualization service in an emergency response. *IEEE Transactions on Services Computing*, 15(2):1251–1264, 2022.

[30] M. E. Kretzschmar, B. Ashby, E. Fearon, C. E. Overton, J. Panovska-Griffiths, L. Pellis, M. Quaife, G. Rozhnova, F. Scarabel, H. B. Stage, B. Swallow, R. N. Thompson, M. J. Tildesley, and D. Villela. Challenges for modelling interventions for future pandemics. *Epidemics*, 38:100546–100546, 2022.

[31] A. Kumpf, J. Stumpfegger, P. F. Hartl, and R. Westermann. Visual analysis of multi-parameter distributions across ensembles of 3d fields. *IEEE Transactions on Visualization and Computer Graphics*, 2021.

[32] A. Kumpf, B. Tost, M. Baumgart, M. Riemer, R. Westermann, and M. Rautenhaus. Visualizing confidence in cluster-based ensemble weather forecast analyses. *IEEE transactions on visualization and computer graphics*, 24(1):109–119, 2017.

[33] L. Liu, M. Mirzargar, R. M. Kirby, R. Whitaker, and D. H. House. Visualizing time-specific hurricane predictions, with uncertainty, from storm path ensembles. In *Computer Graphics Forum*, vol. 34, pp. 371–380. Wiley Online Library, 2015.

[34] G. Marion, L. Hadley, V. Isham, D. Mollison, J. Panovska-Griffiths, L. Pellis, G. S. Tomba, F. Scarabel, B. Swallow, P. Trapman, and D. Villela. Modelling: Understanding pandemics and how to control them. *Epidemics*, 39:100588, 2022. doi: 10.1016/j.epidem.2022.100588

[35] K. Matković, D. Gračanin, M. Jelović, A. Ammer, A. Lež, and H. Hauser. Interactive visual analysis of multiple simulation runs using the simulation model view: Understanding and tuning of an electronic unit injector. *IEEE Transactions on Visualization and Computer Graphics*, 16(6):1449–1457, 2010.

[36] K. Matkovic, D. Gracanin, M. Jelovic, and H. Hauser. Interactive visual steering – rapid visual prototyping of a common rail injection system. *IEEE Transactions on Visualization and Computer Graphics*, 14(6):1699–1706, Nov. 2008. doi: 10.1109/tvcg.2008.145

[37] K. Matkovic, D. Gracanin, B. Klarin, and H. Hauser. Interactive visual analysis of complex scientific data as families of data surfaces. *IEEE Transactions on Visualization and Computer Graphics*, 15(6):1351–1358, 2009.

[38] M. D. Morris. Factorial sampling plans for preliminary computational experiments. *Technometrics*, 33(2):161–174, 1991.

[39] D. Orban, D. F. Keefe, A. Biswas, J. Ahrens, and D. Rogers. Drag and track: A direct manipulation interface for contextualizing data instances within a continuous parameter space. *IEEE transactions on visualization and computer graphics*, 25(1):256–266, 2018.

[40] T. Porphyre, M. Bronsvoort, P. Fox, K. Zarebski, Q. Xia, and S. Gadgil. Scottish covid response consortium (scrc): Eera model overview. 2020.

[41] A. Puy, S. L. Piano, A. Saltelli, and S. A. Levin. sensobol: an r package to compute variance-based sensitivity indices. 2021.

[42] RAMPVIS. RAMPVIS - Visual Analytics Platform. `https://vis.scrc.uk`, (accessed: 31.03.2022).

[43] RAMPVIS. Visualization and visual analytics in support of rapid assistance in modelling the pandemic (ramp). `https://sites.google.com/view/rampvis/`, (accessed: 31.03.2022).

[44] A. Saltelli. Sensitivity analysis for importance assessment. *Risk analysis*, 22(3):579–590, 2002.

[45] A. Saltelli. *Sensitivity analysis in practice : a guide to assessing scientific models*. Wiley, Hoboken, NJ, 2004.

[46] A. Saltelli, K. Aleksankina, W. Becker, P. Fennell, F. Ferretti, N. Holst, S. Li, and Q. Wu. Why so many published sensitivity analyses are false: A systematic review of sensitivity analysis practices. *Environmental modelling 'I&' software : with environment data news*, 114:29–39, 2019.

[47] A. Saltelli, S. Tarantola, and F. Campolongo. Sensitivity analysis as an ingredient of modeling. *Statistical science*, 15(4):377–395, 2000.

[48] A. Saltelli, S. Tarantola, and K. P.-S. Chan. A quantitative model-independent method for global sensitivity analysis of model output. *Technometrics*, 41(1):39–56, 1999.

[49] J. Sanyal, S. Zhang, J. Dyer, A. Mercer, P. Amburn, and R. Moorhead. Noodles: A tool for visualization of numerical weather model ensemble uncertainty. *IEEE transactions on visualization and computer graphics*, 16(6):1421–1430, 2010.

[50] M. Sedlmair, M. Meyer, and T. Munzner. Design study methodology: Reflections from the trenches and the stacks. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2431–2440, 2012. doi: 10.1109/TVCG.2012.213

[51] I. Sobol. Global sensitivity indices for nonlinear mathematical models and their monte carlo estimates. *Mathematics and computers in simulation*, 55(1):271–280, 2001.

[52] I. M. Sobol and S. Kucherenko. Derivative based global sensitivity measures. *Procedia - Social and Behavioral Sciences*, 2(6):7745–7746, 2010. doi: 10.1016/j.sbspro.2010.05.208

[53] B. Swallow, D. A. Rand, and G. Minas. Bayesian inference for stochastic oscillatory systems using the phase-corrected linear noise approximation, 2022. doi: 10.48550/ARXIV.2205.05955

[54] B. Swallow, W. Xiang, and J. Panovska-Griffiths. Tracking the national and regional COVID-19 epidemic status in the UK using directed Principal Component Analysis. *accepted for publication in PTRSA*, 2022. doi: 10.1098/rsta.2021.0302

[55] R. Tavenard, J. Faouzi, G. Vandewiele, F. Divo, G. Androz, C. Holtz, M. Payne, R. Yurchak, M. Russwurm, K. Kolar, and E. Woods. Tslearn, a machine learning toolkit for time series data. *Journal of machine learning research*, 21:1–6, 2020.

[56] T. Torsney-Weir, A. Saad, T. Moller, H.-C. Hege, B. Weber, J.-M. Verbavatz, and S. Bergner. Tuner: Principled parameter finding for image segmentation algorithms using visual response surface exploration. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):1892–1901, 2011.

[57] T. Torsney-Weir, M. Sedlmair, and T. Möller. Sliceplorer: 1D slices for multi-dimensional continuous functions. *Computer Graphics Forum*, 36(3):167–177, June 2017. doi: 10.1111/cgf.13177

[58] J. Wang, S. Hazarika, C. Li, and H.-W. Shen. Visualization and visual analysis of ensemble data: A survey. *IEEE transactions on visualization and computer graphics*, 25(9):2853–2872, 2018.

[59] C. Ware. *Information Visualization: Perception for Design*. Information Visualization: Perception for Design. Elsevier Science, 2013.

[60] R. T. Whitaker, M. Mirzargar, and R. M. Kirby. Contour boxplots: A method for characterizing uncertainty in feature sets from simulation ensembles. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2713–2722, 2013.

[61] W. Xiang and B. Swallow. Multivariate spatio-temporal analysis of the global covid-19 pandemic. *medRxiv*, 2021. doi: 10.1101/2021.02.08.21251339