RESEARCH ARTICLE

The Volatility of Data Space: Topology Oriented Sensitivity Analysis

Jing Du[1]*, Arika Ligmann-Zielinska[2]

1 Department of Construction Science, Texas A&M University, Francis Hall 334, TAMU, College Station, TX 77840, United States of America, **2** Department of Geography, Michigan State University, 673 Auditorium Rd, Room 121, East Lansing, MI 48824, United States of America

* dujing@tamu.edu

Abstract

Despite the difference among specific methods, existing Sensitivity Analysis (SA) technologies are all value-based, that is, the uncertainties in the model input and output are quantified as changes of values. This paradigm provides only limited insight into the nature of models and the modeled systems. In addition to the value of data, a potentially richer information about the model lies in the topological difference between pre-model data space and post-model data space. This paper introduces an innovative SA method called Topology Oriented Sensitivity Analysis, which defines sensitivity as the volatility of data space. It extends SA into a deeper level that lies in the topology of data.

Introduction

Sensitivity Analysis (SA) is "*the study of how uncertainty in the output of a model can be apportioned to different sources of uncertainty in the model input*" [1]. Although it is frequently perceived as an optional step in modeling that can be omitted without a significant loss of information, SA can play a critical role in scientific discovery [2]. It offers a variety of benefits to improve the relevance of modeling to science and technology, including[2–4]:

- identification of critical model factors by quantifying the contribution of each model input variable to the variability of its output, which later allows for efficient allocation of resources for data acquisition;

- legitimate model simplification, which is particularly important when investigating complex systems;

- contribution to theory development by discovering the most accurate representation of the modeled system;

- investigation of deep uncertainties in a variety of systems–a painful but necessary step of scientific discovery;

- comprehensive investigation of model behaviors to provide acceptable policy recommendations in scenario analysis, especially in the absence of outcome scenarios endorsed by stakeholders;

- information provision for stakeholders to develop a shared understanding about the studied problems.

Current SA technologies can be roughly classified into two groups: one-factor-at-a-time (OAT) and global SA. Despite the difference in evaluating the multidimensional model input space, the existing SA technologies are all value-based, where the quantified uncertainties in the model input and output are due to changes of input/output values. For example, a typical OAT examines the change of value of a particular model output when one of the inputs is altered in its value while all other inputs remain unchanged. Another example is a type of global SA based on model output variance decomposition, which results in sensitivity indices that reflect the fraction of output variance (in value) contributable to a certain input.

This value-based paradigm of SA may result in limited insight into the nature of a given model. Fig 1 illustrates a situation when two datasets, A and B, have identical variance, but demonstrate very distinct topological nature of its data points. Specifically, dataset B has a more uniform spatial distribution while A shows a more random pattern. In scenario analysis, if each data point in datasets A and B represents an output scenario, value-based SA would provide little information about the relative "positions" of scenarios in the data space, and may lead to misleading conclusions about policies that satisfy a wide range of future conditions.

This study proposes a new approach to SA called Topology Oriented Sensitivity Analysis (TOSA). We postulate that model sensitivity comprises more than the variance of data, but also the volatility of data space. Consequently, not only does the proposed TOSA capture the change of value of model input and output, but it also provides the means of measuring topological changes of modeled data. It extends SA to a deeper level that lies in the topology of data. Following a brief introduction to the theoretical background, we give a complete description of TOSA. We then demonstrate the utility of TOSA using an agent-based model of shopping behavior as a case study.

Background

Sensitivity Analysis

Based on a comprehensive review of SA literature, Saltelli and Annoni (2010) state that most studies apply SA in an OAT fashion, i.e., changing the value of uncertain factors one-at-a-time while keeping the other factors constant [5]. It has already been found that OAT-SA is justified only for linear models [6, 7]. If the problem is nonlinear, OAT can lead to misleading conclusions [8, 9]. Failure to capture nonlinearities has also been found in regression and correlation
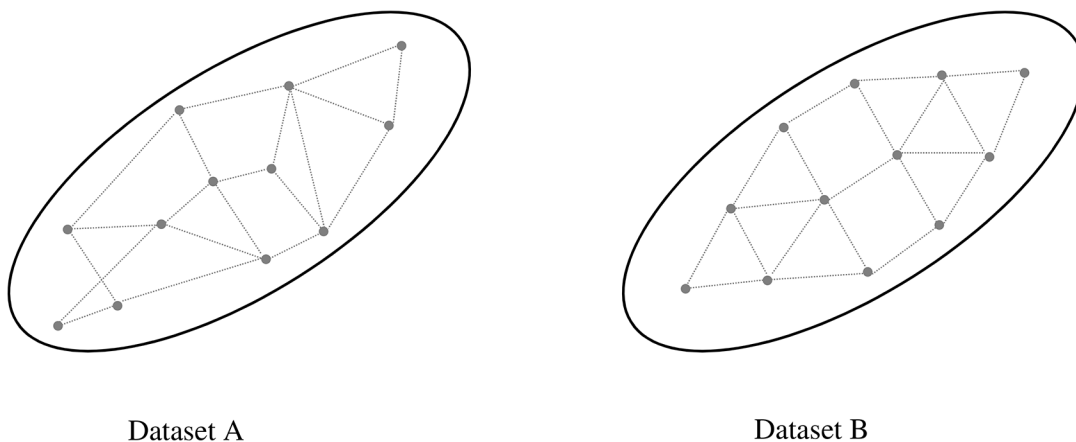
Dataset A   Dataset B

**Fig 1. Two datasets with identical variance are different in nature.**

based SA. Evidence indicates that regression based SA only works for linear models and its effectiveness depends on the goodness of fit [6]. Similarly, correlation measures are not effective at evaluating the sensitivity of complex models since nonlinearities are poorly taken into account in these measures [10].

To address the "curse" of nonlinearity, various remedies have been proposed including the method of Morris and the measure of importance. While the Morris method can account for nonlinearity, it assumes monotonicity, which does not always hold in complex models [10]. Moreover, the Morris method cannot differentiate between the effects caused by model nonlinearity and parameter interactions [10]. The importance measure is also of limited value because it only provides first-order effects (i.e., parameter interactions are not considered) and is very demanding computationally [6].

As a result, variance-based global SA (GSA) has recently received increased attention due to its model independence. In GSA, the unconditional variance of model output is decomposed into terms that account for individual factors plus terms that quantify the interactions among factors [11]. GSA has the capability to account for model nonlinearity and non-monotonicity, regardless of the generic assumptions of the underlying model [6, 10]. In a variety of studies, GSA has proven to do better than the more traditional SA approaches [12]. In the reported literature we identified the extended Fourier Amplitude Sensitivity Test (eFAST) and the method of Sobol as the most popular methods of performing GSA due to their proven track of performance. Sobol's method will be used as a point of departure for the proposed TOSA. In order to explain TOSA, a brief discussion about the modeling theory is necessary.

Redefine Models

Miller and Page [13] have proposed a formal "model of models", which addresses the underlying logic of system modeling. Roughly speaking, the "model of models" reflects the Input-Process-Output (IPO) view of modeling and simulation. The *input* consists of the present states of the system. The model *processes* the input data and generates an *output* in the form of new states of the system. This view puts the model at the center of the simulation process, as shown in Fig 2 left. In this "Model-Oriented" simulation, data is treated mainly as the input and output of a simulation process. The model comes first, then comes the data. An alternative way to
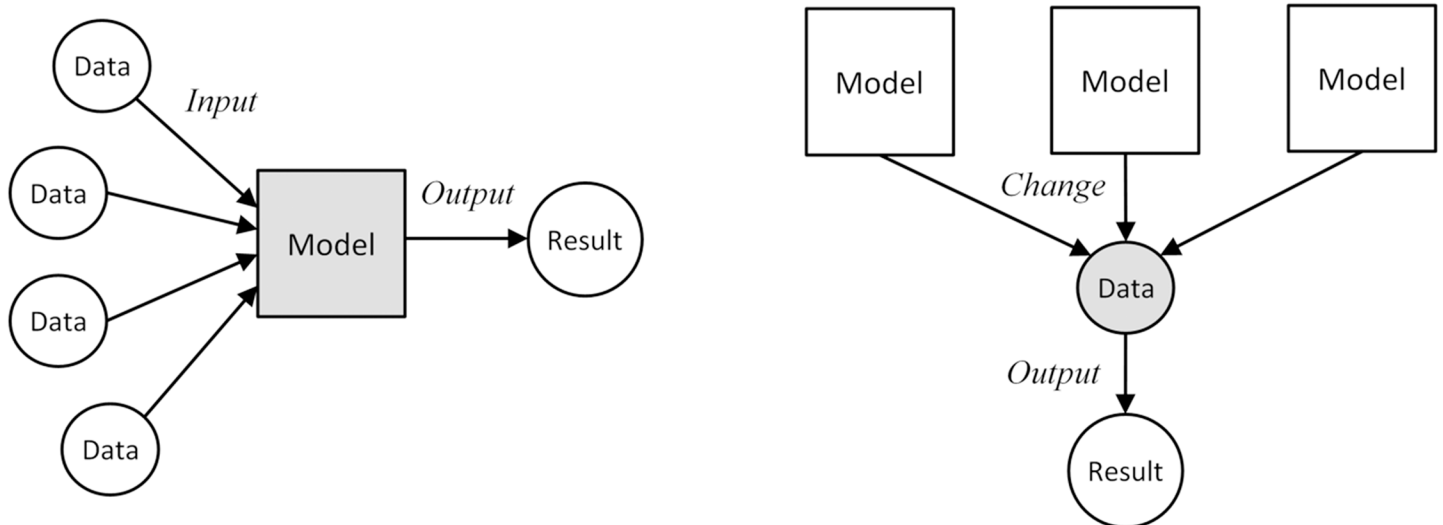
Fig 2. Model-Oriented simulation versus Data-Oriented simulation.

Post-model space

Post-model space
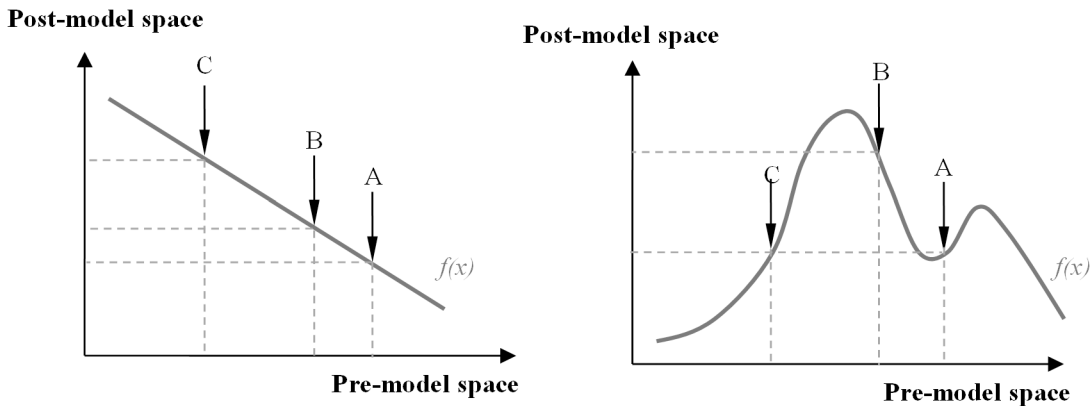
Fig 3. Model is a force that changes data space.

look at the modeling process is the "Data-Oriented" simulation, where data occupies the center stage, and a variety of models are applied to it. Thus, under the "Data-Oriented" approach, the same data is changed by multiple models (Fig 2, right).

The "Data-Oriented" approach to simulation provides additional flexibility to define the modeled data. Data is represented as a space that can be changed by a model. Prior to the model, it is a multidimensional space that contains all the "known" information about the modeled system, After the model, it becomes a multidimensional space that contains all the "resulting states" of the system. A model is a force that changes the configuration of the data space. Fig 3 shows a simple example with one independent and one dependent variable. Assuming that the model is linear, the nearest neighbors A and B in the pre-model space (horizontal) remains the nearest neighbors in the post-model space (vertical; Fig 3 left). When the model is nonlinear, the A and B can end up farther away in the post-model space, with A and C becoming the nearest neighbors (Fig 3 right). Geometrically the space becomes distorted. We will use this data space distortion to redefine sensitivity.

Fig 4 illustrates a case of three dimensions. By traversing through the data space along track A, we capture less variability (uncertainty) than when traversing along track B. If we reduced
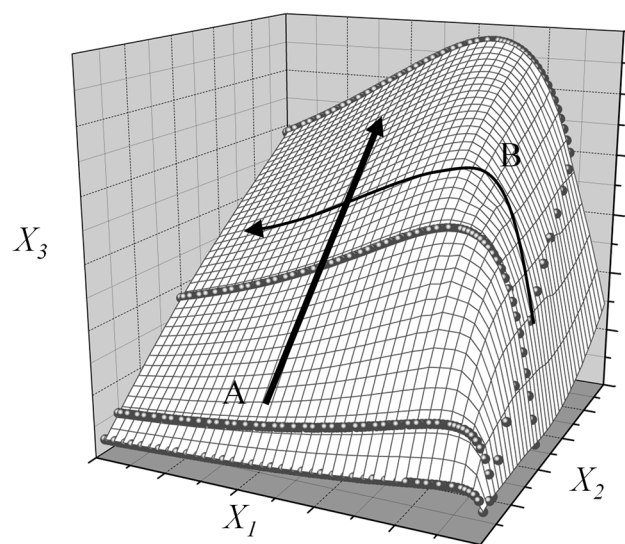
Fig 4. Heterogeneity of the data space.

the dimensionality from 3 to 2 by removing axis X_1, we would not be able to reveal the actual space distortion. We can infer that X_1 contains more information about the model and the data, and thus has a higher level of sensitivity i.e. in the example model outcome is more sensitive to X_1 than X_2 and X_3.

Consequently, sensitivity of a particular variable can be defined as the absolute change of data space when the variable is removed from or added to the model. In the next section, we formulate TOSA to quantify the extent to which the data space distortion can be mitigated or augmented when a certain variable is removed or added.
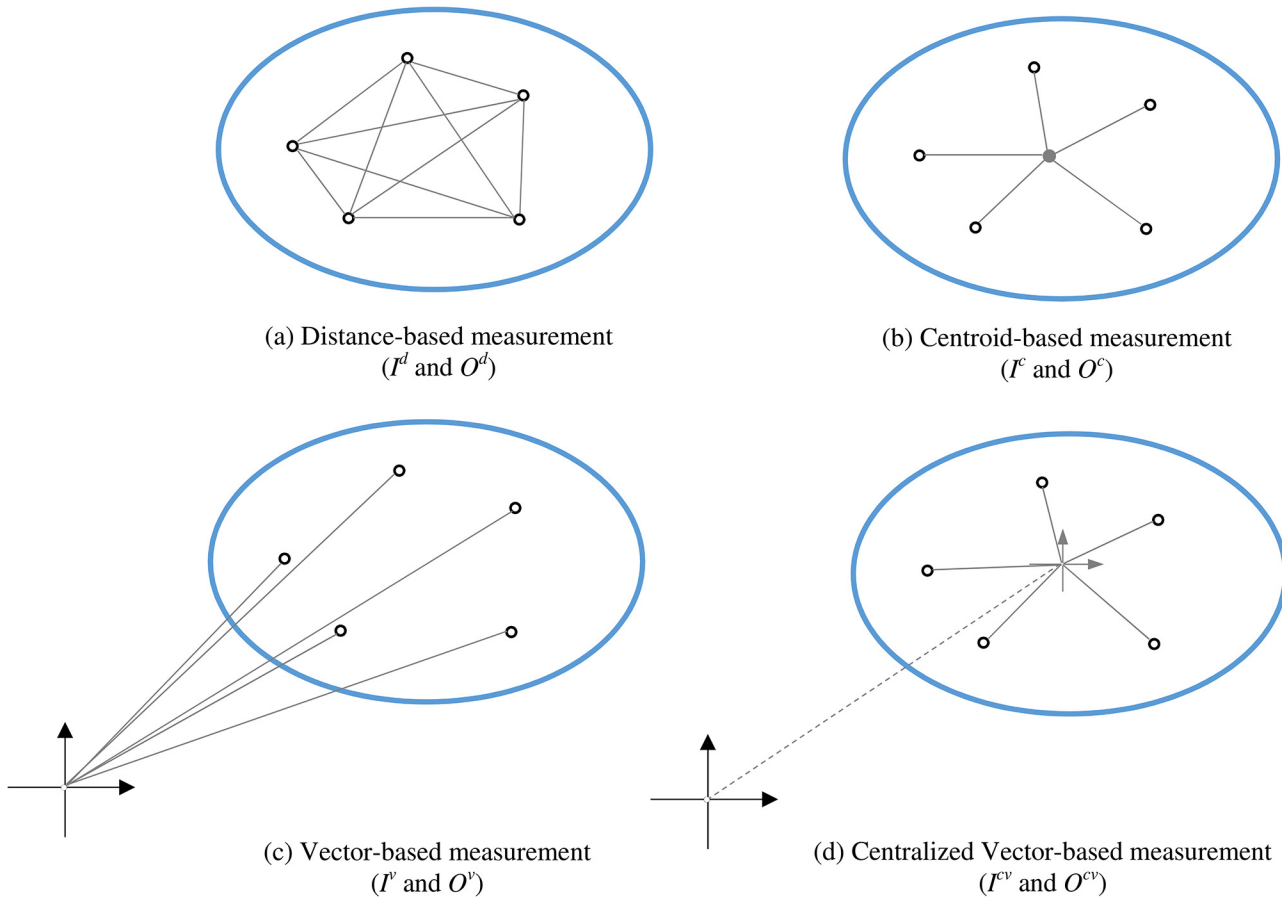
Topology Oriented Sensitivity Analysis (TOSA)

Topological Measurements of Data Space

Define $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)^T$ as the model input, where $n \in [1, \infty)$ is the number of inputs. For any given input $i \in (1,n)$, there is a corresponding $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots x_{im})$, where m is the number of input variables. Note that an input (i) and an input variable ($\mathbf{x_{im}}$) are different concepts, where the former one refers to a data point while the latter one refers to (one of many) variables defining the point. Similarly, define $\mathbf{Y} = (\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n)^T$ as the output, where $\mathbf{y_i} = (y_{i1}, y_{i2}, \ldots y_{im'})$ and m' is the number of output variables. Let $\mathbf{Y} = f(\mathbf{X})$, where f is the model. It is therefore known that $\forall \mathbf{x}_i \in \mathbf{X}: \mathbf{x}_i \in \mathbb{R}^m$, where $\mathbb{R}^m$ is a hyperspace with m dimensions and the ith input of the model is a data point in the hyperspace. Similarly, $\forall \mathbf{y}_i \in \mathbf{Y}: \mathbf{y}_i \in \mathbb{R}^{m'}$, where $\mathbb{R}^{m'}$ is a hyperspace with m' dimensions, and the ith output is a data point $(y_{i1}, y_{i2}, \ldots y_{im'})$. Note that $m \not\equiv m'$, i.e., the number of input variables is not necessarily equal to the number of output variables; but it is assumed that the number of input points is equivalent to the number of output points (both are n). Finally, define $\mathbf{X} \subset \mathbb{R}^m$ as the input data space (i.e., pre-model data space) and $\mathbf{Y} \subset \mathbb{R}^{m'}$ as the output data space (i.e., post-model data space). Formula (1) shows that $\mathbf{X}$ is transitioned to $\mathbf{Y}$ through f.

$$\begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \cdots & \cdots & \vdots \\ x_{i1} & x_{i2} & \cdots & x_{im} \\ \vdots & \cdots & \cdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{pmatrix} \xrightarrow{f} \begin{pmatrix} y_{11} & y_{12} & \cdots & y_{1m'} \\ y_{21} & y_{22} & \cdots & y_{2m'} \\ \vdots & \cdots & \cdots & \vdots \\ y_{i1} & y_{i2} & \cdots & y_{im'} \\ \vdots & \cdots & \cdots & \vdots \\ y_{n1} & y_{n2} & \cdots & y_{nm'} \end{pmatrix} \quad (1)$$

Where $\mathbf{X}$ is distorted by f and its topology is changed. The key step of the proposed SA is to measure the topology of input space ($\mathbf{X}$) and output space ($\mathbf{Y}$). We propose four ways of measuring the change in configuration: distance-based measurement, centroid-based measurement, vector-based measurement, and centralized vector-based measurement (Fig 5).

The distance-based measurement quantifies the distance between each pair of inputs, as illustrated in Fig 5 (A). Suppose the distance between a given pair of inputs $\mathbf{x}_i$ and $\mathbf{x}_j$ (or $\mathbf{y}_i$ and $\mathbf{y}_j$) is D_{ij}, the average D_{ij} indicates the overall configuration of data space topology before (or after) the model takes effect. It is worth noting that $\mathbf{X}$ and $\mathbf{Y}$ are multidimensional spaces and therefore the Mahalanobis distance should be applied instead of the Euclidean distance [14]. Covariance S may exist among variables, exerting unnecessary influence on the calculation of Euclidean distance. The Mahalanobis distance standardizes the distance between any two data points by dividing Euclidean distance by S (covariance matrix), which better captures the relative locations of data points. Following the above notion, the distance between the data points i

(a) Distance-based measurement
(I^d and O^d)

(b) Centroid-based measurement
(I^c and O^c)

(c) Vector-based measurement
(I^v and O^v)

(d) Centralized Vector-based measurement
(I^{cv} and O^{cv})

Fig 5. Four types of topological measurements of data space.

and *j* in the input space and output space are calculated as follows:

$$I_{ij}^d = \sqrt{(\mathbf{x_i} - \mathbf{x_j})^{\mathrm{T}}\mathbf{S_X}^{-1}(\mathbf{x_i} - \mathbf{x_j})} \tag{2}$$

$$O_{ij}^d = \sqrt{(\mathbf{y_i} - \mathbf{y_j})^{\mathrm{T}}\mathbf{S_Y}^{-1}(\mathbf{y_i} - \mathbf{y_j})} \tag{3}$$

And the distance-based measure for the input and output space topology is given by formulas ([4](#)) and ([5](#)), respectively:

$$I^d = \frac{2}{n(n-1)}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}I_{ij}^d \tag{4}$$

$$O^d = \frac{2}{n(n-1)}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}O_{ij}^d \tag{5}$$

Where *n* is the number of inputs (data points or cases) and S_X and S_Y are covariance matrices of input and output data, respectively. Distance-based measurement calculates *n(n-1)/2* pairs of data points. When *n* is big, the computation is nontrivial (e.g. for 100 data points, calculation amounts to 4950 pairs of points).

Data space configuration can also be measured as the distance of each data point to the centroid of the space, as illustrated in Fig 5B. Centroids of input and output spaces are given by:

$$\mathbf{x}_c = \frac{1}{n} \times \sum_{i=1}^{n} \mathbf{x}_i, \; and \; \mathbf{y}_c = \frac{1}{n} \times \sum_{i=1}^{n} \mathbf{y}_i \tag{6}$$

For any given data point, if its distance to the centroid changes, then a distortion of the data space occurs. Using the Mahalanobis distance, the centroid-based measurement can be calculated as the average sum of squares of the distances to the centroid. Formulas (7) and (8) are used to calculate the distance between data point i and the centroid in the input and output space, respectively, and formulas (9) and (10) give centroid-based topological measurements of the input space and output space.

$$I_{ic}^c = \sqrt{(\mathbf{x}_i - \mathbf{x}_c)^\mathrm{T} \mathbf{S}_X^{-1} (\mathbf{x}_i - \mathbf{x}_c)} \tag{7}$$

$$O_{ic}^c = \sqrt{(\mathbf{y}_i - \mathbf{y}_c)^\mathrm{T} \mathbf{S}_Y^{-1} (\mathbf{y}_i - \mathbf{y}_c)} \tag{8}$$

$$I^c = \frac{1}{n} \sum_{i=1}^{n} I_{ic}^c \tag{9}$$

$$O^c = \frac{1}{n} \sum_{i=1}^{n} O_{ic}^c \tag{10}$$

Both the distance-based measurement and the centroid-based measurement use distance to quantify space topology. A common problem faced by both indices is the deterioration of the distance measure when the dimensionality of data is increasing. Beyer et al. [15] proved that the distance between any pair of data points starts to converge to an identical value when dimensionality reaches a certain point, as few as 10–15 dimensions. In other words, for any data point, with the increase in dimensionality, the distance to the nearest data point approaches the distance to the farthest data point:

$$lim_{d \to \infty} = \frac{dist_{max} - dist_{min}}{dist_{min}} = 0 \tag{11}$$

The above phenomenon means that, if the number of input variables (or output variables) is big enough, I_{ij} and O_{ij} will converge to an identical value. To overcome this limitation, a vector-based measurement is proposed.

As illustrated in Fig 5C, if the angle between two input data points $\mathbf{x}_i$ and $\mathbf{x}_j$ is θ_{ij}, then $\cos(\theta_{ij}) = \frac{\mathbf{x}_i \cdot \mathbf{x}_j}{\|\mathbf{x}_i\| \|\mathbf{x}_j\|}$. Note that θ_{ij} is the Euclidean angle. In order to eliminate the influence of covariance among variables, Mahalanobis angle θ_{ij}^M is used [16]. By aggregating $cos(\theta_{ij}^M)$, the space topology can be quantified. Formulas (12) and (13) show the angles between any pair of data points i and j in the input space and output space. Formulas (14) and (15) give vector-

based topological measurements of the input space and output space.

$$I_{ij}^v = \arccos\left(\frac{\mathbf{x}_i^T \mathbf{S}_X^{-1} \mathbf{x}_j}{\sqrt{\mathbf{x}_i^T \mathbf{S}_X^{-1} \mathbf{x}_i}\sqrt{\mathbf{x}_j^T \mathbf{S}_X^{-1} \mathbf{x}_j}}\right) \tag{12}$$

$$O_{ij}^v = \arccos\left(\frac{\mathbf{y}_i^T \mathbf{S}_Y^{-1} \mathbf{y}_j}{\sqrt{\mathbf{y}_i^T \mathbf{S}_Y^{-1} \mathbf{x}_i}\sqrt{\mathbf{y}_j^T \mathbf{S}_Y^{-1} \mathbf{y}_j}}\right) \tag{13}$$

$$I^v = \frac{2}{n(n-1)}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} I_{ij}^v \tag{14}$$

$$O^v = \frac{2}{n(n-1)}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} O_{ij}^v \tag{15}$$

It is worth noting that if data points are farther from the null vector (the origin of the coordinate), and the variance is very small, the angles between any two data points might be too small to demonstrate any significant change. In this case, data needs to be standardized to its center. Formulas (16) and (17) show the centralized angles between any pair of data points i and j in the input space and output space and formulas (18) and (19) give the centralized vector based topological measurements of the input space and output space.

$$I_{ij}^{cv} = \arccos\left(\frac{(\mathbf{x}_i - \mathbf{x}_c)^T \mathbf{S}_X^{-1}(\mathbf{x}_j - \mathbf{x}_c)}{\sqrt{(\mathbf{x}_i - \mathbf{x}_c)^T \mathbf{S}_X^{-1}(\mathbf{x}_i - \mathbf{x}_c)}\sqrt{(\mathbf{x}_j - \mathbf{x}_c)^T \mathbf{S}_X^{-1}(\mathbf{x}_j - \mathbf{x}_c)}}\right) \tag{16}$$

$$O_{ij}^{cv} = \arccos\left(\frac{(\mathbf{y}_i - \mathbf{y}_c)^T \mathbf{S}_Y^{-1}(\mathbf{y}_j - \mathbf{y}_c)}{\sqrt{(\mathbf{y}_i - \mathbf{y}_c)^T \mathbf{S}_Y^{-1}(\mathbf{y}_i - \mathbf{y}_c)}\sqrt{(\mathbf{y}_j - \mathbf{y}_c)^T \mathbf{S}_Y^{-1}(\mathbf{y}_j - \mathbf{y}_c)}}\right) \tag{17}$$

$$I^{cv} = \frac{2}{n(n-1)}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} I_{ij}^{cv} \tag{18}$$

$$O^{cv} = \frac{2}{n(n-1)}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} O_{ij}^{cv} \tag{19}$$

Given the failure of distance calculation in high-dimensional data, vector based measurements are recommended for complex multidimensional models.

Topology Oriented Sensitivity Indices

For any given topological measurement, the change to the spatial relationship between the ith data point and the jth data point after the model takes effect (T_{ij}) can be given by:

$$T_{ij} = \sqrt{(O_{ij} - I_{ij})^2} \tag{20}$$

Where I_{ij} and O_{ij} represent any of the four proposed topological measurements for input space and output space, respectively. The topological change from the input data space to the

output data space (T_X) can be given by:

$$T_X^c = \frac{1}{n} \sqrt{\sum_{i=1}^{n} (O_{ic}^c - I_{ic}^c)^2} \tag{21}$$

$$T_X^{d,v,cv} = \frac{2}{n(n-1)} \sqrt{\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} (O_{ij}^{d,v,cv} - I_{ij}^{d,v,cv})^2} \tag{22}$$

Where n is the number of data points. T_X indicates the extent to which the input data space is distorted by the model given the full set of input variables, i.e., all variables from the 1st to the mth are considered. On the other hand, we can define T_0 as the topological change when none of the variables are considered. Trivially:

$$T_0 = 0 \tag{23}$$

If removing a particular input variable, for example, the variable i, alters the value of T_X, we can infer that the topological change from the input space to the output space is influenced by i. Let's denote T_i as the topological change after the variable i is added into the model, then:

$$T_i = T_{\sim 1,2,\ldots,i-1,i+1,\ldots,m} \tag{24}$$

where $T_{\sim 1,2,\ldots,i-1,i+1,\ldots,m}$ is the topological change after all variables but variable i are removed from the model. Then total topological change of the data space $T(Y)$ is given by:

$$T(Y) = \sum_{i=1}^{n_i} T_i + \sum_{i_1=1}^{n_{i_1}+n_{i_2}} \sum_{i_2=i_1+1}^{n_{i_1}+n_{i_2}} T_{i_1 i_2} + \cdots + T_X \tag{25}$$

where n_i is the number of data points added into the model when variable i is added. $T(Y)$ is the summation of topological changes when one variable, two variables . . . until all variables are added into the system. Following this notion, the Topology Oriented Sensitivity Index (TOSI) is given by:

$$TS_{i_1,\ldots,i_k} = \frac{T_{i_1,\ldots,i_k}}{T(Y)} \tag{26}$$

And the summation of all the TOSI equals 1:

$$\sum_{k=1}^{m} \sum_{i_1 < \ldots < i_k}^{m} TS_{i_1,\ldots,i_k} = 1 \tag{27}$$

If $k = 1$, then $TS_{i_1,\ldots,i_k}$ is called the Main Topology Oriented Sensitivity Index (MTOSI);

$$TS_i = \frac{T_i}{T(Y)} \tag{28}$$

if k $\geq$ 2, then $TS_{i_1,\ldots,i_k}$ is called the Interaction Topology Oriented Sensitivity Index (ITOSI). The Total Topology Oriented Sensitivity Index (TTOSI) is then defined as:

$$TS_i^{tot} = TS_i + TS_{i,\sim i} = 1 - TS_{\sim i} \tag{29}$$

Where $TS_{i,\sim i}$ is the summation of all the $TS_{i_1,\ldots,i_k}$ that involve the variable i and at least one variable from (1,. . ., i-1, i+1, . . . m); and $TS_{\sim i}$ is the summation of all the $TS_{i,\sim i}$ that do not involve any variable i. Consequently, TS_i^{tot} represents the average topological change in the data space that is contributable to the input variable i through its sole influences and interactions with other variables.

Calculation Procedure

TOSI computation requires a particular experimental design. This section explains the calculation steps. In order to make an easy demonstration, we use an example with three input variables X_1, X_2 and X_3, and two output variables Y_1 and Y_2. The model is represented as f. Ten samples are generated in the simulation. Then we build a joint input-output matrix of five columns and ten rows:

$$
\begin{pmatrix}
x_{11} & x_{12} & x_{13} & y_{11} & y_{12} \\
x_{21} & x_{22} & x_{23} & y_{21} & y_{22} \\
x_{31} & x_{32} & x_{33} & y_{31} & y_{32} \\
x_{41} & x_{42} & x_{43} & y_{41} & y_{42} \\
x_{51} & x_{52} & x_{53} & y_{51} & y_{52} \\
x_{61} & x_{62} & x_{63} & y_{61} & y_{62} \\
x_{71} & x_{72} & x_{73} & y_{71} & y_{72} \\
x_{81} & x_{82} & x_{83} & y_{81} & y_{82} \\
x_{91} & x_{92} & x_{93} & y_{91} & y_{92} \\
x_{101} & x_{102} & x_{103} & y_{101} & y_{102}
\end{pmatrix}
\tag{30}
$$

- Step 1: Generate a list of (n) input vectors (1^*m vectors) using a random number sampling approach. m is the number of input variables and n is the number of generated samples. In our example, there are three input variables X_1, X_2 and X_3, and ten randomly generated samples. The input values are showed in the following 3^*10 matrix, where the elements are the randomly generated numbers.

$$
\begin{pmatrix}
x_{11} & x_{12} & x_{13} \\
x_{21} & x_{22} & x_{23} \\
\vdots & \ddots & \vdots \\
x_{101} & x_{102} & x_{103}
\end{pmatrix}
\tag{31}
$$

- Step 2: Execute the model n times with the generated input vectors. Each input vector is a line in the above matrix. In our example, for the first vector (x_{11}, x_{12}, x_{13}), we obtain $(y_{11}, y_{12}) = f(x_{11}, x_{12}, x_{13})$. This step is repeated until the last execution $(y_{101}, y_{102}) = f(x_{101}, x_{102}, x_{103})$.

- Step 3: Calculate T_X following formulas (21) and (22). Matrix (30) is used to calculate T_X. Note there are four types of T_X specified by equations (2) through (19).

- Step 4: Calculate the average value of the n samples given any input variable X_i, denoted as $\bar{x}_i$, and remove input vectors except $(x_1, x_2, \ldots, \hat{x}_i, \ldots, x_m)$, where $\hat{x}_i \approx \bar{x}_i$. Similarly, remove output vectors except $(y_1, y_2, \ldots, \hat{y}_i, \ldots, y_{m'})$, where $(x_1, x_2, \ldots, \hat{x}_i, \ldots, x_m) \xrightarrow{f} (y_1, y_2, \ldots, \hat{y}_i, \ldots, y_{m'})$. For example, we calculate the average value of all samples given X_1. Suppose the average value is close to x_{21}, x_{51} and x_{71}, then we remove all input vectors and corresponding output vectors other than the second, the fifth

and the seventh vectors. By doing it, the randomness of X_1 has been ruled out (only leaving the average value), and thus its impact on the outputs is eliminated. The new matrix is:

$$\begin{pmatrix} x_{21} & x_{22} & x_{23} & y_{21} & y_{22} \\ x_{51} & x_{52} & x_{53} & y_{51} & y_{52} \\ x_{71} & x_{72} & x_{73} & y_{71} & y_{72} \end{pmatrix} \tag{32}$$

- Step 5: Calculate $T_{1,2,\ldots,i-1,i+1,\ldots,m} = T_{\sim i}$ following formulas (21) and (22). In our example, matrix (32) is used to calculate the new T_X. The results is denoted as $T_{\sim 1}$ to illustrate that X_1 has been removed. After the calculation of this step is complete, the removed vectors are put back to matrix (32).

- Step 6: Repeat steps 4 and 5 until all x_i are consecutively removed, and the corresponding $T_{\sim i}$'s are calculated. In our example, $T_{\sim 1}$, $T_{\sim 2}$ and $T_{\sim 3}$ are calculated, which means X_1, X_2 and X_3 are removed separately from matrix (30).

- Step 7: For any input vector $(x_1, x_2, \ldots, \hat{x}_i, \ldots, x_m)$, calculate the average value of input variable x_j denoted as $\bar{x}_j$, where $i \neq j$, and remove input vectors except $(x_1, x_2, \ldots, \hat{x}_i, \ldots, \hat{x}_j, \ldots, x_m)$, where $\hat{x}_i \approx \bar{x}_i$ and $\hat{x}_j \approx \bar{x}_j$. Similarly, remove output vectors except $(y_1, y_2, \ldots, \hat{y}_i, \ldots, \hat{y}_j, \ldots, y_{m'})$, where $(x_1, x_2, \ldots, \hat{x}_i, \ldots, \hat{x}_j, \ldots, x_m) \xrightarrow{f} (y_1, y_2, \ldots, \hat{y}_i, \ldots, \hat{y}_j, \ldots, y_{m'})$. In our example, suppose X_1 has been removed (matrix (32)). Then the average value of the remaining samples given X_2 is calculated, which is close to x_{22} and x_{72}. Thus we need to remove the fifth input vector and the corresponding output vector. The new matrix is:

$$\begin{pmatrix} x_{21} & x_{22} & x_{23} & y_{21} & y_{22} \\ x_{71} & x_{72} & x_{73} & y_{71} & y_{72} \end{pmatrix} \tag{33}$$

Where the impact of both X1 and X2 on outputs have been eliminated. It is therefore possible to evaluate the importance of these input variables by "removing" them from the model.

- Step 8: Calculate $T_{1,2,\ldots,i-1,i+1,\ldots,j-1,\ldots,j+1,\ldots,m} = T_{\sim ij}$ following formulas (21) and (22). In our case, matrix (33) is used to calculate the new T_X, which is denoted as $T_{\sim 12}$, to illustrate that X_1 and X_2 have been removed.

- Step 9: Repeat steps 7 and 8 until $x_1, \ldots, x_k$ ($2 \leq k \leq m$) are removed, and all the corresponding $T_{\sim 1,2,\ldots,k}$ ($2 \leq k \leq m$) are calculated. In our case, the calculated T_X's include $T_{\sim 1}, T_{\sim 2}, T_{\sim 3}$, $T_{\sim 12}, T_{\sim 13}$ and $T_{\sim 23}$.

- Step 10: Calculate MTOSI, ITOSI and TTOSI following formulas (25) through (29).

Case Study

The URBAN model

We demonstrate the proposed TOSA using a model called URBAN, an agent-based model (ABM) developed to investigate how individual shopping travel behaviors reshape urban configuration [17]. Two types of agents are modeled in URBAN: *Store* and *Person*. As illustrated in Fig 6, the red dots represent stores, and the green squares are people. When an individual
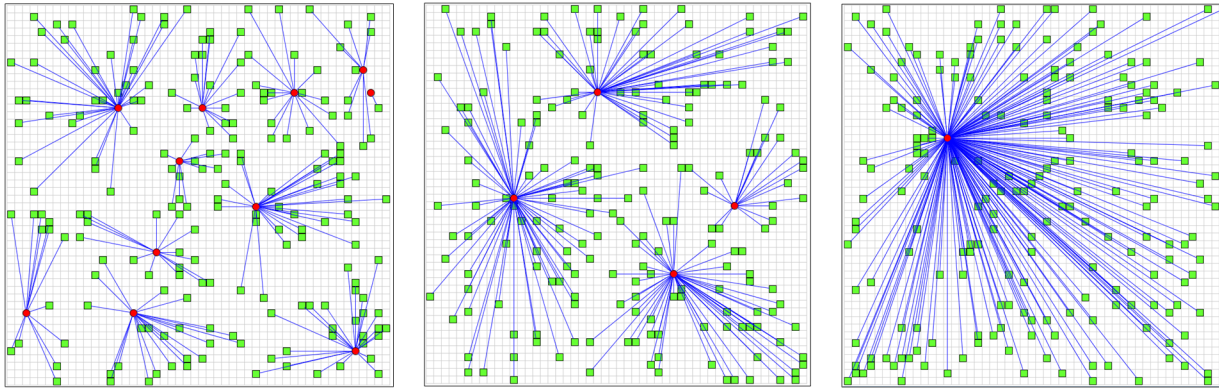
Fig 6. The URBAN model—sample results.

decides to go shopping at a particular store, a connection is established (represented by a blue line in Fig 6). The go/no go decision of a person depends upon three factors: (1) the utility of a store (including the size of the store, the selection of goods, etc.); (2) store accessibility (including store distance and traveling expenses); and (3) the socioeconomic status of the individual. If a store is visited by more people, it maintains higher profits and grows in size, which in turn becomes more attractive to other potential shoppers; in contrast, if a store is visited less frequently, it starts to lose profit, shrinks in size, and ultimately runs out of business (it is removed from the ABM).

Table 1 summarizes the main input and output variables of URBAN. Note that, for the demonstration purposes, the initial number of stores, the initial number of people, and the size of urban space are held fixed in the computation of TOSI's. Four critical input variables are examined in the simulations: preference for store utility (α), preference for accessibility (β), dollar amount of purchase, and store operating cost per day. Note that α and β are directly related to individual decision making: if the value of α is bigger for an agent, it means she/he cares more about the store utility; and if β is bigger, it means that accessibility is more important to the agent. The other two input variables, *dollar amount of purchase* and *store operating cost per day*, influence the dynamics of the store status: if a store has a higher daily operating cost (in percent of the cash reserve), it may run out of business faster if not enough income is obtained. On the other hand, if dollar amount of purchase is higher, the stores may be able to maintain the income. The output variables of interest are the final number of stores, the total walking distance and the total driving distance. These output variables are used to determine the final urban configuration.

The dollar amount of purchase and the store operating cost per day were assumed to follow a normal distribution (Table 1); socioeconomic status, measured as household income, was assumed to follow a Pareto distribution.

Table 1. Input and output variables of the URBAN model.

Symbol	Variable name	Values	Symbol	Variable name	Values
			X6	Initial # of stores	20
	Input		X7	Initial # of people	200
X1	Preference for store	0~6, step 0.25			
X2	Preference for travel	0~6, step 0.25		*Output*	
X3	$ amount of purchase	N (100, 20)	Y1	Final number of stores	N/A
X4	Store operating cost per day	N (0.01,0.001)	Y2	Total walking distance	N/A
X5	Urban space size	50*50 gird where a grid cell equals 100^2 square feet	Y3	Total driving distance	N/A

TOSI

All the four types of TOSI's were calculated, including the distance-based indices, the centroid-based indices, the vector-based indices, and the centralized vector based indices. The a and β coefficients were changed in a OAT (one at a time) manner from 0 to 6, at the step 0.25. Because Probability Density Functions (PDFs) were used to describe certain input variables, for a given combination of a and β, the simulation was repeated 30 times to generate a statistically sound results. As a result, a total of 18,750 simulations were performed. The input data space is therefore a seven-dimensional space with 18,750 data points, and the output space is a three-dimensional space with the same number of data points.

The centroid based indices were first calculated. Fifteen terms were calculated including all the main and interaction topology-oriented indices, each involving 18,750 matrix calculations. Given the difficulty of calculation, a Visual Basic (VB) program was developed to facilitate the calculation. The selection of VB is a practical choice. Many business practices rely on Microsoft Excel. It would be easier for nonacademic users to apply our tools in Excel as a VB application (VBA). For academic users, we also provided the tools programed in SAS and R [18], which more efficiently handle matrix data structures.

Table 2 summarizes the results. After normalizing the result to 100% (% TS_i and % TS_i^{tot}), we identified the *preference for the store utility* (X1) and *preference for accessibility* (X2) to be

Table 2. Centroid based TOSI's.

No.	Combination			
1	X1	X2	X3	X4
2		X2	X3	X4
3	X1		X3	X4
4	X1	X2		X4
5	X1	X2	X3	
6			X3	X4
7		X2		X4
8		X2	X3	
9	X1			X4
10	X1		X3	
11	X1	X2		
12	X1			
13		X2		
14			X3	
15				X4
	X1	X2	X3	X4
$T_{i,\sim i}$	0.3396	0.3318	0.2283	0.2310
TS_i	0.2239	0.2211	0.0936	0.0943
TS_i^{tot}	0.4427	0.4326	0.2977	0.3012
% TS_i	35.4%	34.9%	14.8%	14.9%
% TS_i^{tot}	30.0%	29.3%	20.2%	20.4%

X1: preference for the store utility (α)
X2: preference for accessibility (β)
X3: dollar amount of purchase
X4: store operating cost per day.

doi:10.1371/journal.pone.0137591.t002

Table 3. Calculation results.

MTOSI	X1	X2	X3	X4
Distance based	28.9%	28.1%	21.2%	21.8%
Centroid based	35.4%	34.9%	14.8%	14.9%
Vector based	35.0%	31.8%	18.9%	14.3%
Centralized Vector based	29.6%	29.4%	20.6%	20.4%
TTOSI	**X1**	**X2**	**X3**	**X4**
Distance based	28.3%	27.8%	21.7%	22.2%
Centroid based	30.0%	29.3%	20.2%	20.4%
Vector based	31.3%	29.9%	21.5%	17.3%
Centralized Vector based	28.7%	28.7%	21.3%	21.4%

more influential on the topological change than the *dollar amount of purchase* (X3) and *store operating cost* (X4).

We also calculated the other three types of TOSI. Compared to the centroid-based indices, the calculation was more difficult, as each of the 15 terms involves 175,771,875 matrix calculations. A set of VB programs was developed to facilitate the calculation. The results are summarized in Table 3. Although different in exact values, all TOSI's suggest stronger influences of *preference for the store utility* (X1) and *preference for accessibility* (X2).

In order to evaluate the difference between TOSA and traditional SA, we performed an OAT SA on the same data. To be noted, unlike TOSA that gives a single sensitivity index in spite of the number of output variables, the OAT SA needs to be done for each of the output variables. Fig 7 shows that the three output variables—final number of stores (store), the total walking distance (walk) and the total driving distance (drive)—are highly correlated. As a result, we only present the results of OAT SA for output "drive".
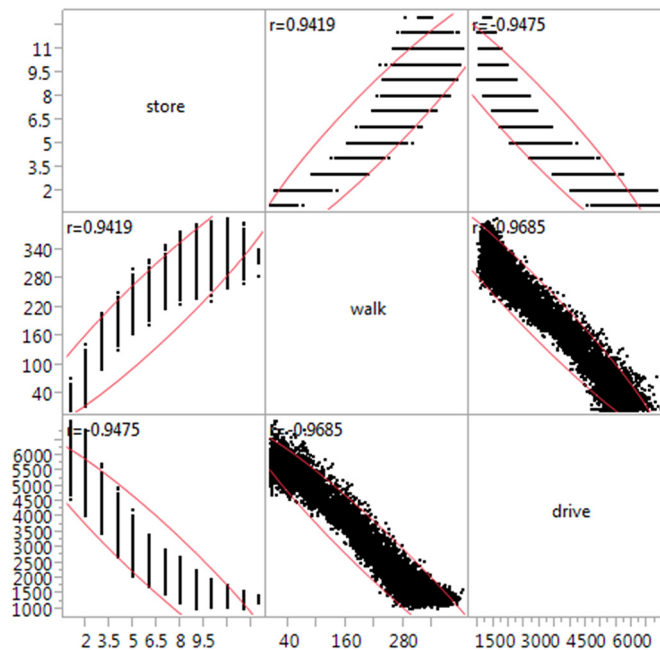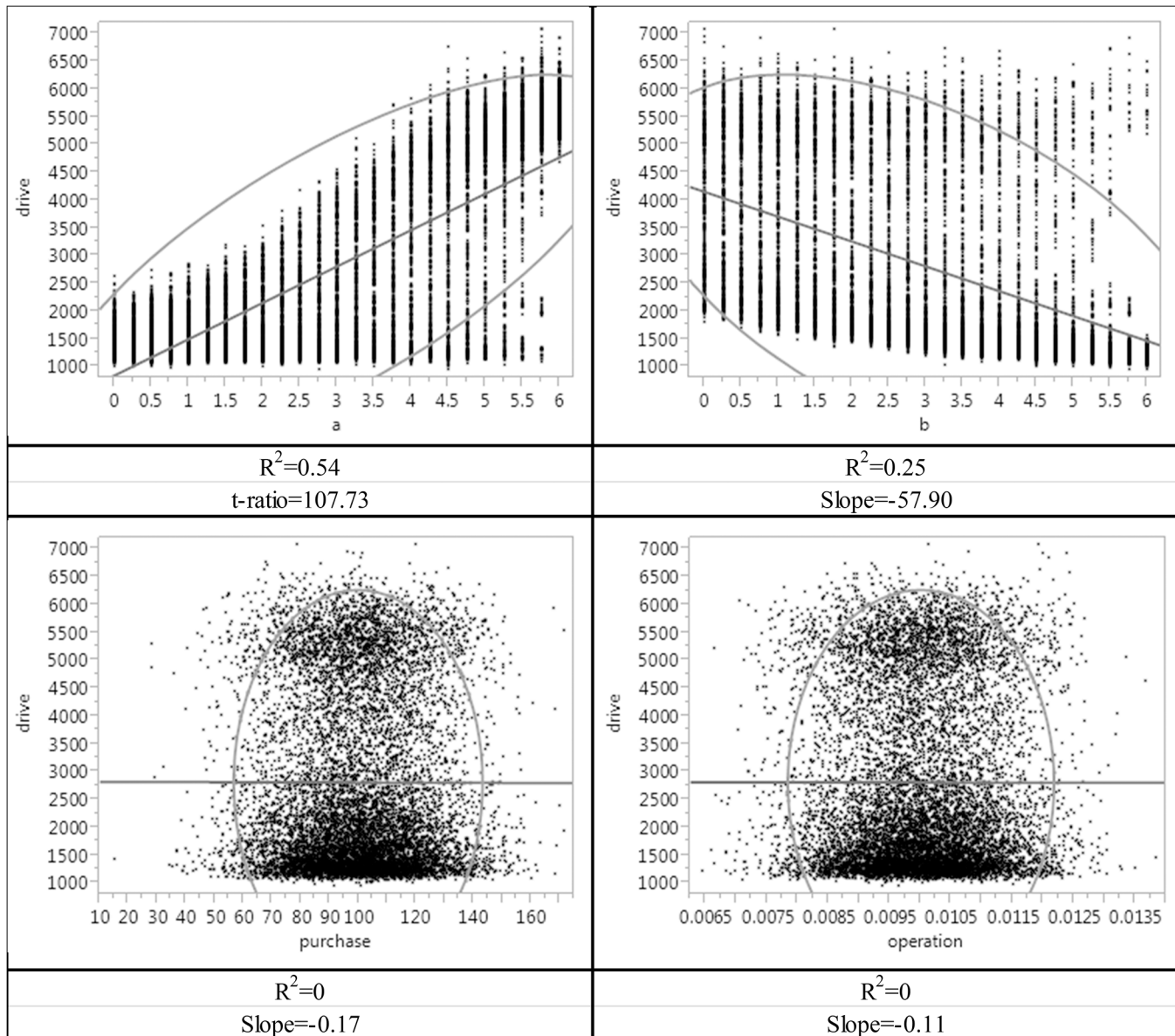

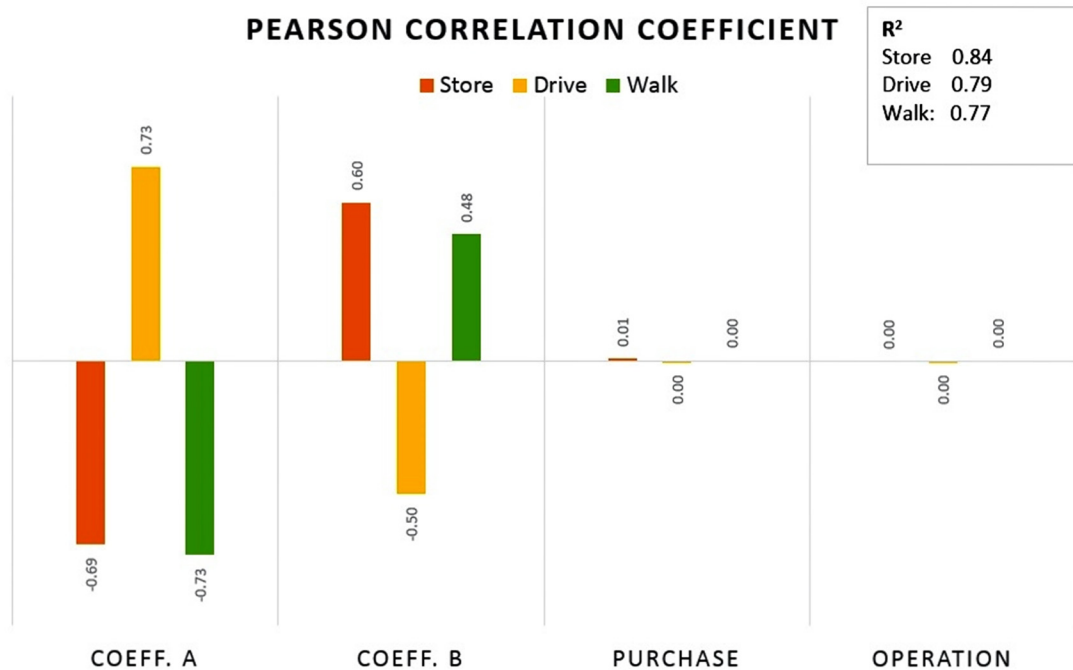**Fig 7. Correlation of the three output variables.**

Fig 8. OAT sensitivity analysis on total driving distance.

doi:10.1371/journal.pone.0137591.g008

To obtain comparable measures, we summarized the results of OAT using linear regression. As shown in Fig 8, coefficients α and β are more influential on the output variable "drive", while "purchase" and "operation" do not show any observable impact on this output. If using the t-ratio of the linear regression (which is the standardized coefficient of input variable in a regression analysis) to evaluate the degree of impacts, α outperforms β. In other words, according to the OAT SA, *preference for the store utility* (X1) and *preference for accessibility* (X2) can significantly affect the total driving distance, with the former exhibiting stronger influence. However, the OAT SA does not support the impact of *dollar amount of purchase* (X3) or *store operating cost* (X4) on the total driving distance.

We also used Pearson's correlation coefficient (r)—one of the simplest measures of global SA [19]. Fig 9 depicts the results of the analysis, together with the coefficients of determination.

PEARSON CORRELATION COEFFICIENT

■ Store   ■ Drive   ■ Walk

R^2	
Store	0.84
Drive	0.79
Walk:	0.77

Fig 9. Pearson correlation coefficient calculated for the three output variables.

The following observations can be made. Given all three output measures, the model behaves fairly linearly as about 80% of the influence can be attributed to individual inputs ($R^2 \sim 0.8$). The influence of inputs is similar for "store" and "walk" with the coefficients α and β mostly affecting the variability of the outputs. Note that where the impact of α is negative, the impact of β is positive. For 'drive', the impact is essentially inverted. In absolute terms, however, α and β coefficients are the most influential on all three output variables, with α slightly dominating β.

After comparing the results of the three methods, we found that TOSA, OAT SA and Global SA (based on the correlation coefficient), can lead to quite different conclusions. Although all three methods unanimously suggest that *reference for the store utility* (X1) is the most influential factor and *preference for accessibility* (X2) is the second, TOSA also indicates the importance of *dollar amount of purchase* (X3) or *store operating cost* (X4). It is because the existence of *dollar amount of purchase* (X3) or *store operating cost* (X4) can affect the level of topological change of the data, which is the foundation of TOSA calculation. Practically, it means that although altering *dollar amount of purchase* (X3) or *store operating cost* (X4) will not change the expected number of stores, or the expected walk/driving distance, the specific combination of the three urban setup indicators has actually changed. For example, Fig 10 illustrates the simulation result when *dollar amount of purchase* (X3) and *store operating cost* (X4) are changing, while the values of *reference for the store utility* (X1) and *preference for accessibility* (X2) are both fixed to 3.0. As shown, although the expected outcome is always the same (it is the centroid of the 3D scatter plot), it actually represents very different futures–without changing X1 and X2, the maximum driving distance is still 22.5% longer than the minimum driving distance in only 16 simulations. The output data space is still very volatile, suggesting potentially distinct futures. Traditional value-based SA methods fail to capture this nuance because this piece of subtle information is hidden behind data topology rather than the mean value or its variance. Clearly, TOSA provides a different yet complementary angle to interpret model sensitivity.

A further comparison has highlighted the differences between TOSA and traditional SA. Traditional value-based SA methods rely heavily on summary statistics (variance, mean) which may fail when dealing with non-monotonic and non-additive models [20, 21]. While useful in many circumstances, the correlation coefficient (and its derivatives like Spearman, standardized regression coefficients etc.) is derived from a linear model, and does not provide any information on the interaction effects among inputs which, in our example, account for about 20% of model variability. Published research suggests that some of the inputs may not be influential singly, but may substantially affect the variability of results when evaluated in combination with other inputs, contributing to higher-order effects [22]. In addition, the measures of sensitivity may be different for different output variables (in our case, for each input variable, there is a separate sensitivity index for "store", "walk" and "drive" respectively), posing a challenge to a modeler when deciding on factor fixing and model simplification. As an alternative, we could use ANOVA-like measures of sensitivity (aka variance-based global SA), but these methods require a quasi-random experimental design, which may be inappropriate when other types of post-processing analyses (based on parametric statistics) are also employed to the output data.
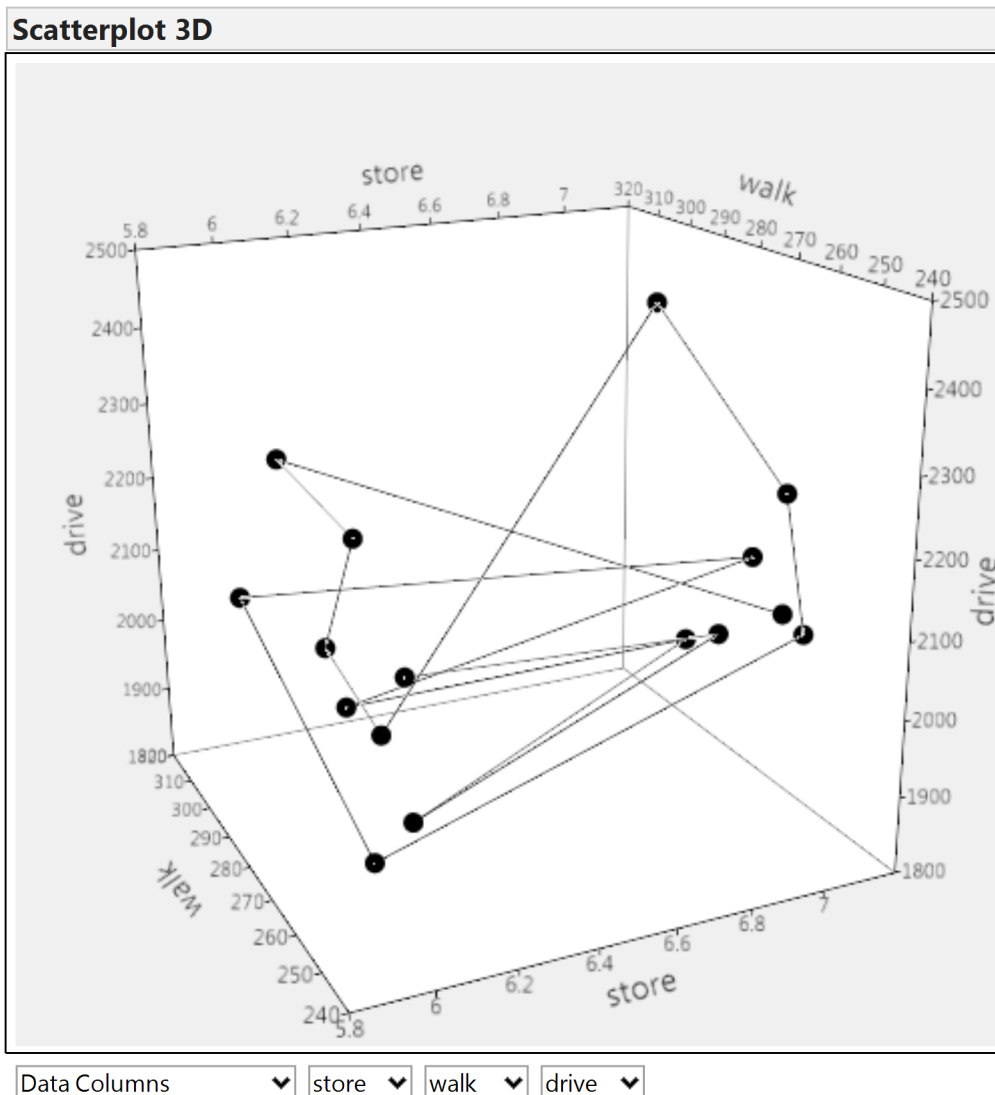
Fig 10. The volatile outcome when the values of X1 and X2 are fixed to 3.0.
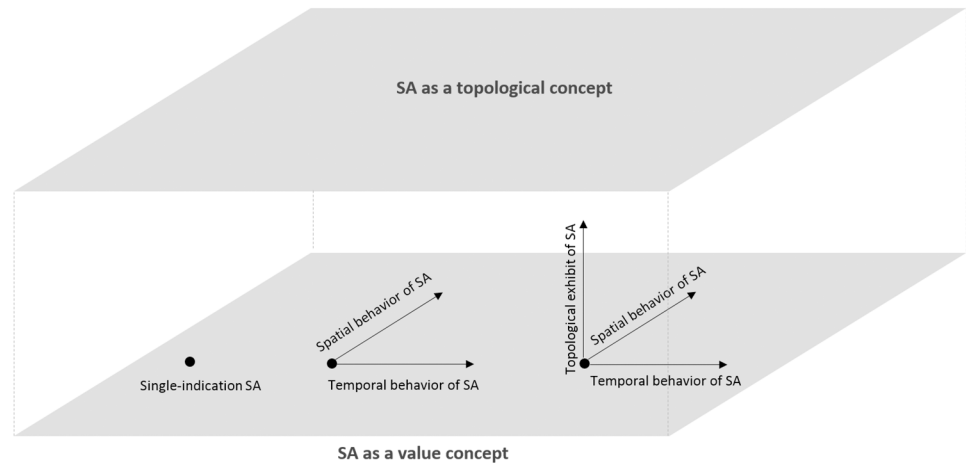
Fig 11. Evolution of SA methodology.

Discussion

Model sensitivity has to be evaluated in relation to the specific context of a modeling study. In general however, the concept of model sensitivity relates to the relationships between model input and output uncertainties. Model SA should therefore measure the degree of change after the model takes effect. Most SA methods define uncertainty using statistical denotations, either as the absolute change in values, or as their variance. We argue that topology, which is not present in the commonly-utilized approaches to SA, may provide additional useful information on model behavior and the complexity of relationships between its inputs and outputs.

The proposed TOSA attempts to quantify the topological difference between model input data space and its output data space. It builds on a view of data-oriented simulation (Fig 2), in which models are external mechanisms that distort the data space. A cross-paradigm property of models is their ability to alter the data space topology. In this light, sensitivity is an indicator of the volatility of data space: if adding a dimension (an input variable) strengthens the volatility of the modeled data space, the outcome is sensitive to the added dimension. This new angle of SA is a promising avenue for model exploration and evaluation.

Recent studies demonstrate how SA can be applied to evaluate the temporal [2, 23] and spatial [24–26] complexities of models. Temporal SA explores the regions in the time series of sensitivities where a particular input dominates the others. Spatial SA investigates the spatial heterogeneities of models, especially in the geographic space [24]. Compared to the traditional SA, in which only the final model sensitivities are of interest, temporal SA and spatial SA examine the behaviors of a model over time and space. SA has been therefore extended from a scalar (single-indication) type of analysis to a time series or a layer that contains spatially differentiated information (Fig 11). Following the previous works, TOSA adds a new dimension to comprehensive model evaluation, where SA is applied as a topological concept, other than a value concept (Fig 11). As a consequence, a new school of SA methods may emerge, promoting our understanding of models and the modeled systems. We hypothesize that TOSA may be especially useful in model-based scenario analysis, contributing to more solid understanding of factors that are critical in identifying similar output scenarios.

The proposed methodology is obviously in its infancy. Future work will focus on three aspects of TOSA development. First, using a number of case studies, we plan to compare and contrast TOSA with the most common value-based approaches (most notably, variance-based SA). Second, we will identify conditions in which the use the simpler centroid-based TOSA

metrics is insufficient. Third, the temporal complexity of TOSA could be as high as $O(m!)$ times $O(n^2)$. In the case study, the longest experiment took nearly nine hours to finish (on a dual processor desktop computer with 32 GB RAM, running Windows 7). We argue, however, that modelers should be first concerned with getting the right answer, and then focus on reducing the processing time of the evaluation procedure. We do not expect TOSA to be done in real time. In many applications, post-modeling analyses are performed to obtain more knowledge about the studied systems, rather than assisting in real time decision making. Even when the real time analysis is required, some technologies may be leveraged to expedite the calculation, such as cloud computing. In the future, we plan to optimize the execution to provide a new tool that can be applied to diverse complex modeling applications.

Conclusions

SA is undoubtedly a critical component in shaping our understanding of modeled systems. Current SA approaches quantify sensitivities on the basis of change in values. This paper proposes a different approach to evaluating model behavior—an approach based on topology that represents the connectivity of multidimensional data points. Datasets with identical statistical features (e.g., variance) may be differently spaced, resulting in diverse topological structures. As a result, two factors of identical influence on output variability (measured using the common SA approaches), may have a different effect when evaluated within the topological space.

The proposed Topology-Oriented Sensitivity Analysis captures the topological difference between the pre-model data space and post-model data space. It defines sensitivities of a particular variable as the contributed marginal and interactive topological changes when this variable is added to the model. When the data space demonstrates more volatility after a variable is added, it suggests a high level of model sensitivity to this variable. Measuring volatility of data space is not a trivial task, since the dimensionality of data space keeps changing during TOSA calculation (removing/adding dimensions). Therefore, as a benchmark, we calculate the ratio of the output space topology to the input space topology.

Topology-based sensitivity analysis introduces an alternative way of looking at a model and its data. It introduces new opportunities for investigating the hidden but potentially critical characteristics of modeled systems. More efforts are therefore urged to extend this paradigm of model evaluation.

Supporting Information

S1 Text. VBA code for Distance based TOSI.
(TXT)

S2 Text. VBA code for Centroid based TOSI.
(TXT)

S3 Text. VBA code for Vector based TOSI.
(TXT)

S4 Text. VBA code for Centralized Vector based TOSI.
(TXT)

S5 Text. SAS code for Distance based TOSI.
(TXT)

S6 Text. SAS code for Centroid based TOSI.
(TXT)

S7 Text. SAS code for Vector based TOSI.
(TXT)

S8 Text. SAS code for Centralized Vector based TOSI.
(TXT)

S9 Text. R code for all four TOSI's.
(TXT)

Acknowledgments

Author Contributions

Conceived and designed the experiments: JD ALZ. Performed the experiments: JD. Analyzed the data: JD ALZ. Contributed reagents/materials/analysis tools: JD. Wrote the paper: JD ALZ. Coding: JD.

References

1. Saltelli A, Ratto M, Andres T, Corporation E. Global sensitivity analysis: the primer: Wiley Online Library; 2008.

2. Ligmann-Zielinska A, Sun L. Applying time-dependent variance-based global sensitivity analysis to represent the dynamics of an agent-based model of land use change. International Journal of Geographical Information Science. 2010;99999(1:):1–22.

3. Lempert R, Popper S, Bankes S. Shaping the next one hundred years: New methods for quantitative, long-term policy analysis: Rand Corporation; 2003.

4. Ligmann-Zielinska A, Kramer DB, Cheruvelil KS, Soranno PA. Using Uncertainty and Sensitivity Analyses in Socioecological Agent-Based Models to Improve Their Analytical Performance and Policy Relevance. PloS one. 2014; 9(10):e109779. doi: 10.1371/journal.pone.0109779 PMID: 25340764

5. Saltelli A, Annoni P. How to avoid a perfunctory sensitivity analysis. Environmental Modelling & Software. 2010; 25(12):1508–17.

6. Lilburne L, Tarantola S. Sensitivity analysis of spatial models. International Journal of Geographical Information Science. 2009; 23(2):151–68.

7. Saltelli A, Ratto M, Tarantola S, Campolongo F. Sensitivity analysis practices: Strategies for model-based inference. Reliability Engineering & System Safety. 2006; 91(10):1109–25.

8. Thogmartin WE. Sensitivity analysis of North American bird population estimates. Ecological Modelling. 2010; 221(2):173–7.

9. Varella H, Guérif M, Buis S. Global sensitivity analysis measures the quality of parameter estimation: the case of soil parameters and a crop model. Environmental Modelling & Software. 2010; 25(3):310–9.

10. Maier H, Dandy G, Norton J, Croke B. A Comparison of Sensitivity Analysis Techniques for Complex Models for Environmental Management. Integrated Assessment. 2005:2533–9.

11. Chen W, Jin R, Sudjianto A. Analytical variance-based global sensitivity analysis in simulation-based design under uncertainty. Journal of mechanical design. 2005; 127:875.

12. Sallaberry CJ, Helton JC, editors. An Introduction to Complete Variance Decomposition. 24th Conference and Exposition on Structural Dynamics 2006 (IMAC—XXIV); 2006; St Louis, Missouri, USA.: Curran Associates, Inc.

13. Miller J, Page S. Complex adaptive systems: An introduction to computational models of social life: Princeton Univ Pr; 2007.

14. Mahalanobis PC, editor On the generalized distance in statistics. Proceedings of the National Institute of Sciences of India 2 (1); 1936.

15. Beyer K, Goldstein J, Ramakrishnan R, Shaft U. When is "nearest neighbor" meaningful? Database Theory—ICDT'99: Springer; 1999. p. 217–35.

16. Raich A, Çinar A. Diagnosis of process disturbances by statistical distance and angle measures. Computers & chemical engineering. 1997; 21(6):661–73.

17. Du J, Wang Q. Exploring Reciprocal Influence between Individual Shopping Travel and Urban Form: Agent-Based Modeling Approach. Journal of Urban Planning and Development. 2011; 137(4):390–401.

18. Team RC. R: A Language and Environment for Statistical Computing 2015. Available from: http://www.R-project.org.

19. Saltelli A, Chan K, Scott EM. Sensitivity Analysis. Chichester, England: Wiley-Interscience; 2000. 475 p.

20. Borgonovo E, Castaings W, Tarantola S. Model emulation and moment-independent sensitivity analysis: An application to environmental modelling. Environmental Modelling & Software. 2012; 34(0):105–15. doi: http://dx.doi.org/10.1016/j.envsoft.2011.06.006.

21. Borgonovo E. Measuring uncertainty importance: investigation and comparison of alternative approaches. Risk Analysis. 2006; 26(5):1349–61. PMID: 17054536

22. Ligmann-Zielinska A, Sun L. Applying Time Dependent Variance-Based Global Sensitivity Analysis to Represent the Dynamics of an Agent-Based Model of Land Use Change. International Journal of Geographical Information Science. 2010; 24(12):1829–50.

23. Saltelli A, Tarantola S, Chan K. A role for sensitivity analysis in presenting the results from MCDA studies to decision makers. Journal of Multi-Criteria Decision Analysis. 1999; 8(3):139–45.

24. Ligmann-Zielinska A. Spatially-explicit sensitivity analysis of an agent-based model of land use change. International Journal of Geographical Information Science. 2013; 27(9):1764–81.

25. Ligmann-Zielinska A, Jankowski P. Spatially-explicit integrated uncertainty and sensitivity analysis of criteria weights in multicriteria land suitability evaluation. Environmental Modelling & Software. 2014.

26. Marrel A, Iooss B, Jullien M, Laurent B, Volkova E. Global sensitivity analysis for models with spatially dependent outputs. Environmetrics. 2011; 22(3):383–97.