Global Sensitivity Analysis of Input variables for a Train Accident Risk Model

Monika Reif

Institute of Applied Mathematics and Physics, Zurich University of Applied Sciences, Switzerland Department, E-mail: monika.reif@zhaw.ch

Joanna Weng

Institute of Applied Mathematics and Physics, Zurich University of Applied Sciences, Switzerland Department, E-mail: joanna.weng@zhaw.ch

Christoph Zaugg

Institute of Applied Mathematics and Physics, Zurich University of Applied Sciences, Switzerland Department, E-mail: christoph.zaugg@zhaw.ch

Safe transportation of hazardous materials by rail is an important issue in Switzerland. This study analyzes an existing model for the risk of transport of hazardous materials via Swiss railways, in collaboration with the Swiss Federal Office for the Environment. The model is the basis for the risk calculation of hazards for persons for all railway transports of hazardous materials in Switzerland and is published by the Swiss Federal Office of Transport. It includes 155 input variables estimated with different uncertainties. The objective of this study is to determine which input variables possess the strongest influence on the model output (the risk) and should therefore be determined with higher accuracy.

To achieve this objective, different sensitivity analysis methods as suggested by Borgonovo are compared. The risk model is implemented in Maple and the Sobol decomposition is used for a global sensitivity analysis of the input variables. The Sobol method is a variance-based sensitivity analysis that decomposes the variance of the output of the model into contributions due to input variables or sets of input variables. The Sobol indices are calculated analytically by evaluating various integrals in the decomposition. In addition, the stability of the method is investigated by using different ranges of the input variables. As a first cross check, the partial derivatives of all input variables are calculated for the same model. As a second cross check, an independent analysis in Matlab is carried out, based on Monte Carlo simulation of the input variables within their uncertainty range.

The results are stable and consistent among all methods and will be used by the Swiss Federal Office for the Environment to optimize the estimation of the input variables of this risk model.

Keywords: Variance-Based Sensitivity Analysis, Uncertainty Quantification, Sobol Indices, Sobol Decomposition, Probabilistic Risk Assessment, Transportation of Hazardous Materials by Rail

1. Introduction

The Swiss Federal Office of Transport (FOT) developed a model for the risk of the transport of hazardous materials via Swiss railways. In this quantitative risk assessment, many input variables are used with different quality (accuracy of the estimate) and influence on the risk. To improve the quality of the results and thus the assessment of the risk, mathematical methods to determine the most important variables are required. In this paper several methods for a sensitivity analysis are implemented and compared in order to determine the influence of the variables on the model output (i.e. the risk). The analysis is carried out in behalf of the Swiss Federal Office for the Environment (FOEN). Based on the results, future efforts of the variable assessments by the Swiss Federal Office of Transport and the Swiss Federal Office for the Environment can be focused on the most relevant variables.

First, the risk model is described followed by an overview of the methods used to quantify the relevance of input variables. Then the technical implementations of the risk model in the Maple and Matlab programs are outlined. Finally, the results are presented and discussed.

2. Structure of the Risk Model

Proceedings of the 30th European Safety and Reliability Conference and the 15th Probabilistic Safety Assessment and Management Conference
Edited by Piero Baraldi, Francesco Di Maio and Enrico Zio
Published by Research Publishing, Singapore.
ISBN: 978-981-14-8593-0; doi:10.3850/978-981-14-8593-0

Hazardous materials are assigned to one of three categories, based on similar chemical properties. The categories are represented by gasoline, propane and chlorine.

In this paper the risk for the chlorine category is analyzed. The output of the model is the annual risk per a 100 m track section. The structure of the model consists of three parts.

In the first part the location-specific release frequency for materials in the chlorine category is calculated taking into account several variables like the quantity of dangerous material, railway topology, etc.

The release frequency is used in the second part as initiating event for the event tree which models different events that influence the accident progression. Each branch of the tree is referred to as an accident scenario.

The third part quantifies the consequences, i.e. how many people are killed per accident scenario in the immediate vicinity of the railway track or in a passenger train nearby. The output of the model (i.e. the total risk) is the sum of all accident scenarios times casualties per scenario. Details are described in BAV (2015).

3. Methods to Quantify the Relevance of Input Variables

After a search in the literature, (Borgonovo (2017) for an extensive list of references) the following two methods have been chosen to quantify the relevance of the real valued input variables. We first describe briefly the nature of both methods and clarify the reasons for choosing them. Details of the implementation will be given thereafter.

Method 1 relies on I. M. Sobol's decomposition (Sobol (1993)) which represents a model output as a sum of effects of all orders. It consists of a global method based on an analysis of variance, also well suited to detect possible interactions between various input variables. We will show that in our case, first order effects already explain more than 99% of the variance of the model output. Therefore, we will only focus on these effects in the following. This focusing reduces the computational effort, such that the main effects as well as their variances can be calculated symbolically with a Computer Algebra System (CAS).

Method 2 uses partial derivatives (Borgonovo (2017)) and has the advantage, that the values of the corresponding sensitivity indices can be calculated with relatively little computing effort. Since sensitivity indices based on partial derivatives may be negative, this method offers an independent check of the implementations. The disadvantage of this method is that it is only local, i.e. the values of the sensitivity indices may heavily depend on the choice of the base case defined below.

The results of these methods are confirmed by Monte Carlo simulations.

3.1 *Notation and Terminology*

A number n of real valued input variables x_i is collectively denoted by:

$$x = (x_1, .., x_n) \tag{1}$$

The x_i are considered to be random variables, uniformly distributed on closed intervals. A base case B consists of specifying a fixed real value B_i for each input variable:

$$B = (B_1, .., B_n) \tag{2}$$

A positive real valued parameter α controls the boundary of the closed interval associated to the input variable x_i. The left end of the interval is:

$$a_i = (1 - \alpha)B_i \tag{3}$$

The corresponding right end is given by:

$$b_i = (1 + \alpha)B_i \tag{4}$$

The Cartesian product of all closed intervals is a cuboid $\mathrm{Q}(n)$ with the base case B at its center:

$$\mathrm{Q}(n) = [a_1, b_1] \times .. \times [a_n, b_n] \tag{5}$$

A real valued function $F(x)$ depends on n real input variables x_i and models the output.

3.2 *Sobol Decomposition*

The Sobol decomposition represents the model output $F(x)$ as a sum of all possible effects. In our case this sum is well approximated by the effects up to order one:

$$F(x) \approx f_0 + \sum_{i=1}^{n} f_i(x_i) \tag{6}$$

The constant f_0 equals the expected value of the model output and is given by the following n-dimensional integral:

$$f_0 = \int_{\mathrm{Q}(n)} dx\, F(x) \tag{7}$$

An effect of order one is a real function depending only on one input variable x_i. To determine this

quantity, we have to calculate the $(n-1)$-dimensional integral:

$$f_i(x_i) = \int_{Q(n-1)} dx_{\sim i}\, F(x) - f_0 \qquad (8)$$

Integration in Eq. (8) is taken with respect to all input variables except x_i. The variance of the model output is the n-dimensional integral:

$$V_0 = \int_{\mathrm{Q}(n)} dx\, F^2(x) - f_0^2 \qquad (9)$$

Finally, the variance V_i of the effect $f_i(x_i)$ is calculated as:

$$V_i = \int_{a_i}^{b_i} dx_i\, f_i^2(x_i) \qquad (10)$$

The variance of the model output equals the sum of variances of all effects. In our case the following approximation holds:

$$V_0 \approx \sum_{i=1}^{n} V_i \qquad (11)$$

The sensitivity with respect to input variables x_i is quantified by the Sobol index:

$$S_i = \frac{V_i}{V_0} \qquad (12)$$

Table 1 summarizes the computational complexity in terms of the numbers of one-dimensional integrations to be performed:

Table 1. Computational complexity

Quantity	Meaning	Number of integrations
f_0	Expected value	n
V_0	Model variance	n
$f_i(x_i)$	Effect of x_i	$n(n-1)$
V_i	Variance	n

3.3 *Partial Derivatives*

The differential dx_i of the input variable x_i is assumed to be proportional to the corresponding base case B_i:

$$dx_i = \beta B_i \qquad (13)$$

The constant of proportionality β is dimensionless and equals the same numeric value for all input variables x_i. The index of sensitivity G_i is defined to be:

$$G_i = \frac{\partial_i F(B) B_i}{\sum_{i=1}^{n} \partial_i F(B) B_i} \qquad (14)$$

These indices are normalized:

$$\sum_{i=1}^{n} G_i = 1 \qquad (15)$$

In contrast to the Sobol indices, their values can also be negative, i.e. an increase in the input variable x_i may result in a decrease of the model output $F(x)$. In order to compare the ranks of the most sensitive input variables for both methods, the quantities $|G_i|$ are used.

3.4 *Monte Carlo Simulation*

The two methods chosen are validated by Monte Carlo simulation, used to investigate the reaction of a model to randomly generated input values.

In general, the method is carried out according to the following three steps:

- *N* sets of inputs are generated randomly and thus describe *N* possible base cases.
- For each base case an evaluation is performed, i.e. the output of the model is determined depending on the randomly generated inputs.
- The outputs of the base cases are evaluated by calculating the mean value, the distribution, the minimum or maximum value, etc. of the output.

In this context it is important that suitable algorithms are used for the generation of the random numbers. This is especially important in the case of several input variables, since the entire range of possible combinations of variables must be covered by the (quasi) random numbers (Sobol (1990)).

4 Implementation

The risk model introduced in chapter 2 consists of a real valued function $F(x)$ modeling the output, which depends on $n = 155$ real input variables x_i, i.e. the number of casualties per track section per year.

All input variables may be classified as follows:

- Release frequency (initiating event): The first part of the risk model contains roughly 80 variables and involves knowledge concerning the railway operation, e.g. mass of the hazardous materials transported, number of trains or railway network topology.

- Event tree (branches): The second part of the risk model contains about 10 variables describing various circumstances of the accident progression and thus the accident scenario.
- Consequences: The third part of the risk model contains the remaining variables modelling the number of casualties among humans affected in case of an accident, depending on the nature of the track section, e.g. population density, distance from the accident.

4.1 *Sobol Decomposition and Partial Derivatives*

Sobol decomposition and partial derivatives are implemented in Maple, taking into account the variables for the release frequency, the events in the event tree and the consequences (casualties). For each input variable a range of variation is defined according to Eq. 3 and Eq. 4 using different values for α, see table 4.

4.2 *Monte Carlo Simulation*

To verify the results, the risk model is additionally implemented in Matlab. The structure of the event tree is programmed and the model output for the mean value of the base cases is calculated. The same procedure is repeated for the minimum value and the maximum value of the base cases considered. The same boundaries are used as in the Sobol decomposition.

This serves to determine the risk of the base case as well as the risk arising from the marginal values to evaluate the Maple and Matlab implementations.

In the first simulation the expected value of the risk is determined in order to compare it with the result for the Sobol decomposition and for the partial derivatives. A point in the cuboid $\mathrm{Q}(n)$ is randomly chosen according to the uniform distributions within the defined intervals.

In the second simulation all input variables are assigned to the base case B. Only one variable is chosen randomly within the boundaries. The other variables are each frozen at the value of the base case. Afterwards the variable of the highest influence on the output of the model is determined based on the variance.

All simulations are performed with 20,000 replications each. For each input variable a range of variation is defined based on the maximum and minimum values given in the report. For some parameters, the range has been extended according to the assumptions in the risk model.

5 Results

5.1 *Results of the Sobol Decomposition*

The Sobol decomposition provides a ranking for the influence of the various variables on the determination of personal risk arising from chlorine transportation accidents.

For this purpose, the sensitivity S_i of the input variables x_i is quantified by the share of the variance of the effect of these variables in the model variance V_0. The ten model variables with the highest values for sensitivity S_i are listed in Table 2.

Table 2. Variables with the highest sensitivity indices in % of the share for Sobol decomposition (cf. Eq.(12)).

Rank	Description	S_i (%)
1	Correction factor for improved tank wagons	18.1
2	Correction factor for hot box and brake detection systems	18.1
3	Correction factor for train velocity	18.1
4	Release rate per full tank wagon	18.1
5	Daytime wind	8.6
6	Day/night	3.7
7	Evacuation possibility	3.2
8	Quantity released	2.4
9	Average lethality outdoors, 200-2'500m	1.7
10	Nighttime wind	1.3

The four largest values for sensitivities S_i have variables whose values are used in the first part of the model, covering the calculation of the probability of a train accident with release of hazardous material.

The subsequent variables with high sensitivities describe variables of the event tree such as day/night, wind day or night, evacuation possibility (Table 2), release quantity and one of the values of lethality. These variables contribute to the calculation of the accident progression and the consequences of the accident.

The sensitivities of the first four or five values differ significantly from the following ones. The values of the input variables of the first five ranks together already explain a share of more than 80% of the variance of the entire model. Input variables

with ranks six to ten explain about another 12% of the variance.

From rank 11 onwards, the contribution of the variables to the overall variance is less than 1% and the 145 variables in ranks 11 to 155 together explain less than 8% of the overall variance.

5.2 *Results of the Partial Derivatives*

The second method of evaluating the influence of the variables on the basis of the partial derivatives reveals an identical ranking, as can be seen in Table 3.

Table 3. Variables with the highest sensitivity indices in % of the share for Sobol decomposition (cf. Eq. (12)) and partial derivatives (cf. Eq. (14)).

Rank	Description	$\|G_i\|$ (%)	S_i (%)
1	Correction factor for improved tank wagons	12.7	18.1
2	Correction factor for hot box and brake detection systems	12.7	18.1
3	Correction factor for train velocity	12.7	18.1
4	Release rate per full tank wagon	12.7	18.1
5	Daytime wind	8.7	8.6
6	Day/night	5.7	3.7
7	Evacuation possibility	5.3	3.2
8	Quantity released	4.6	2.4
9	Average lethality outdoors, 200-2'500m	3.9	1.7
10	Nighttime wind	3.4	1.3

The ranking is based on the absolute value of the sensitivity index $|G_i|$, since the sensitivity index can also take on negative values.

The negative G_i values can be explained by the fact that in the model, the event tree is mapped as it appears in the original risk model for determining personal risks (BAV (2015)). In most cases, the event tree of the risk model is designed in such a way that the event that leads to a higher extent of loss is represented in the upper path of a branch. This is not the case at some branches, and therefore negative partial derivatives also arise at these points, since the risk decreases rather than increases with larger values of the model variables, e.g. higher wind intensity reduces chlorine concentration and leads to a lower number of casualties.

5.3 *Consistency checks and comparison with the Monte Carlo method*

For the Sobol decomposition and the partial derivatives, the expected value of the risk is $2.57 \cdot 10^{-5}$ casualties per track element and year.

The results of the Matlab cross check are:

- Risk with mean values: $2.58 \cdot 10^{-5}$ casualties per track element and year
- Risk with minimum values: $2.33 \cdot 10^{-14}$ casualties per track element and year
- Risk with maximum values: $1.62 \cdot 10^{-3}$ casualties per track element and year

These results coincide with the results of the Maple implementation.

Afterwards the results of a Monte Carlo simulation with (quasi) random uniformly distributed values for all variables reveals a similar value for the expected value. The calculated mean value of the 20,000 replications provides an estimate for the expected value of $2.57 \cdot 10^{-5}$ casualties per track element and year. This value corresponds to the value of the model for the two variants described above.

The second simulation is used to verify the variable sensitivity derived by the methods described above. Thereby only one parameter is varied as mentioned above and the absolute differences and the variance of the calculated risk are determined as a function of the variable.

As a result of this simulation it is found that factors contributing to the calculation of the probability of an accident provide the greatest variances. This corresponds to the observation made for the two methods selected, Sobol decomposition and partial derivatives.

5.4 *Evaluation of the Stability of the Ranking (Sobol)*

In order to check the stability of the ranking of the two proposed methods, the base case is chosen as the mean value in each case and the interval of the variables is varied by changing parameter α.

For the results given so far, an interval of the variables in a range of - 5% and + 5% around the base case has been defined. To test the stability of the ranking, this interval is extended to ± 10%, ± 15% and ± 20%.

Table 4. Variation of the interval (α parameter).

α	Expected value	Variance (model)	Explained variance (1st order effects)
5%	$2.6 \cdot 10^{-5}$	$3.04 \cdot 10^{-12}$	99.8%

10%	$2.6 \cdot 10^{-5}$	$1.22 \cdot 10^{-11}$	99.2%
15%	$2.6 \cdot 10^{-5}$	$2.78 \cdot 10^{-11}$	98.1%
20%	$2.6 \cdot 10^{-5}$	$5.01 \cdot 10^{-11}$	96.7%

If the base case is enlarged in this way, the expected value remains stable, the variances in the model increase and the variances explained by first-order effects decrease slightly (higher-order variances become more significant), see Table 4.

Table 5 contains the significance indices of the top ten of the Sobol decomposition depending on the different intervals for the variables. With increasing interval size (α values), the values of the sensitivity indices decrease. This means that the variables with higher ranks have an increasing influence on the calculated risk at larger intervals. However, the ranking of the top ten remains stable.

Table 5. Highest sensitivity indices in % of the share for Sobol decomposition (cf. Eq.(12)) depending on parameter α.

Rank	Description	S_i (%) α=5%	S_i (%) α=10%	S_i (%) α=15%	S_i (%) α=20%
1	Correction factor for improved tank wagons	18.1	18.0	17.8	17.6
2	Correction factor for hot box and brake detection systems	18.1	18.0	17.8	17.6
3	Correction factor for train velocity	18.1	18.0	17.8	17.6
4	Release rate per full tank wagon	18.1	18.0	17.8	17.6
5	Daytime wind	8.60	8.55	8.46	8.34
6	Day/night	3.67	3.64	3.61	3.55
7	Evacuation possibility	3.23	3.21	3.17	3.13
8	Quantity released	2.39	2.37	2.35	2.31
9	Average lethality outdoors, 200-2'500m	1.74	1.73	1.71	1.69
10	Nighttime wind	1.28	1.28	1.26	1.25

6 Summary

It is shown that both the Sobol decomposition and the calculation of partial derivatives are suitable for the analysis of a risk model, which bases on a calculation of the probability of an accident with the modelling of the resulting consequences using an event tree.

Our approach to analyze sensitivity is based on two features of the model under consideration:

1. The structure of the model: The output is a sum of products of powers in the input variables.
2. An approximation property: The sum of variances due to first order effects almost exhausts the variance of the model output.

In combination, these two features enable a symbolic calculation of the relevant sensitivity indices in an efficient way. Accuracy of the results of our approach is validated by two alternative methods involving partial derivatives and Monte Carlo simulations respectively.

Our approach is feasible for models sharing these two features and containing less than 300 input variables. Limiting the number of input variables is necessary if using a personal computer.

The result of this analysis shows which variables have the greatest influence on the output. Based on this result, the next step could be to determine the most relevant variables more precisely and/or to simplify the risk model by reducing the number of variables. This might imply a change in the sensitivity ranking.

Furthermore, the applicability of these sensitivity methods to other risk models basing on other approaches, e.g. Bayesian nets, could be examined.

References

FOT (Swiss Federal Office of Transport) (2015). Risks to the population transporting dangerous materials by rail. In German: Risiken für die Bevölkerung beim Transport gefährlicher Güter auf der Bahn, Bundesamt für Verkehr (BAV).

Borgonovo, E (2017) Sensitivity Analysis. In: Springer International Series in Operations Research & Management Science (Volume 251). Editor: Hillier, F.S.

Sobol, I.M. (1993) Sensitivity Estimates for Nonlinear Mathematical Models, MMCE, VOL.1, NO 4, pp. 407-414.

Sobol, I.M. (1990) Quasi-Monte Carlo methods, Progress in Nuclear Energy, Vol. 24, Issues 1–3, pp. 55-61.