SENSITIVITY ANALYSIS WITH CORRELATED INPUTS – AN ENVIRONMENTAL RISK ASSESSMENT EXAMPLE

Srikanta Mishra
INTERA Inc.
9111A Research Blvd
Austin, TX 78758 USA

ABSTRACT

Crystal Ball® calculates sensitivities by computing rank correlation coefficients between model inputs (assumptions) and outputs (forecasts) – an approach that is known to provide inaccurate results for correlated assumptions. This paper describes the Partial Correlation Coefficient (*PCC*) concept for sensitivity analysis of probabilistic models with correlated inputs. *PCC*s quantify the strength of a linear relationship between input-output pairs after eliminating the linear influence of other input variables, and can be readily calculated from the input-input correlation matrix and the input-output correlation vector. The methodology is illustrated using an analytical model of environmental health risk arising from groundwater-borne radionuclide migration from a nuclear waste repository.

1 INTRODUCTION

Sensitivity analysis, in its simplest sense, involves quantification of the change in model output corresponding to a change in one or more of the model inputs. In the context of probabilistic models, however, sensitivity analysis is generally taken to imply identification of input parameters that have the greatest influence on the spread (variance) of model results (Helton, 1993). This is also referred to as global sensitivity or uncertainty importance analysis to distinguish it from the classical sensitivity measures obtained as partial derivatives of the output with respect to inputs of interest (Saltelli et al., 2000).

The contribution to output uncertainty (variance) by an input is a function of both the uncertainty of the input variable and the sensitivity of the output to that particular input. In general, input variables identified as important in global sensitivity analysis have both characteristics; they demonstrate significant variance and are characterized by large sensitivity coefficients. Conversely, variables which do not show up as important per these metrics are either restricted to a small range in the probabilistic analysis, and/or are variables to which the model outcome does not have a high sensitivity.

A commonly-used measure of input-output sensitivity or uncertainty importance is Spearman's rank correlation coefficient, *RCC*, defined as (Helton et al., 1991):

$$RCC[y,x_k]=\frac{\sum_k (x_k-\bar{x})(y_k-\bar{y})}{\left[\sum_k (x_k-\bar{x})^2 \sum_k (y_k-\bar{y})^2\right]^{1/2}} \tag{1}$$

where x is the input of interest, y is the output, the overbar symbol denotes the sample mean and k is an index for the samples (realizations). The *RCC* provides a measure of the degree to which the input variable of interest and the output can change together. It quantifies the strength of linear and monotonic association between the input-output pair – with the rank transformation facilitating a linearization of any underlying non-linear trends (Helton, 1993). Positive values of the *RCC* imply that an increase in the input corresponds to an increase in the output, with negative values implying the reverse situation. The larger the absolute value of the *RCC*, the stronger the relationship between the input-output pair. The *RCC* is also the primary measure used by Crystal Ball for ranking the most important variables in a probabilistic model.

When a linear additive input-output model is built with uncorrelated inputs, the goodness-of-fit of the model can be expressed as (Draper and Smith, 1981):

$$R^2=\sum_j RCC^2[y,x_j] \tag{2}$$

where R^2, the coefficient of determination, denotes the fractional variance in y explained by the model. Thus, the term $RCC^2[y,x_j]$ can be interpreted as the fractional variance in y explained by the j-th independent variable. As can be easily ascertained, both Eq. (1) and Eq. (2) lead to the same order of importance for the uncertain inputs. It should be pointed out that Crystal Ball uses Eq. (2) to determine the fractional contribution to output variance by the uncertain inputs as an alternative measure of uncertainty importance.

When some of the input variables are correlated, the goodness-of-fit of the input-output model can no longer be expressed via a simple linear sum as in Eq. (2), but must also include terms reflecting the covariance of the correlated inputs. In such situations, it becomes difficult to assign a unique component of the output variance to each of the uncertain inputs. Crystal Ball recognizes this limitation, and recommends in the User's Manual that the importance ranking on the basis of *RCC*s, as depicted in the Sensitivity Chart, should be carefully used when inputs are correlated.

2 PARTIAL CORRELATION CONCEPT

The partial correlation coefficient, *PCC*, measures the correlation between the output and the selected input variable after the linear influence of the other variables have been eliminated (Draper and Smith, 1981). The partial rank correlation coefficient, *PRCC*, is the corresponding measure when input-output relationships are built using the ranks of the variables to linearize the relation. With little loss of generality, we will use *PRCC*s in the following discussion – with the understanding that the input-output pair of interest has already been rank transformed.

Let y denote the output variable and x_j, $j = 1,\ldots, n$, denote the uncertain inputs – some of which are correlated. In order to determine the *PRCC* between y and the p-th uncertain input, x_p, we first build a linear regression model between y and all the other uncertain inputs, viz:

$$\hat{y} = b_o + \sum_{j \neq p} b_j x_j \tag{3}$$

where b denotes a regression coefficient and the 'hat' signifies a regression-fitted variable. Next, a linear regression model is built between x_p and all the other uncertain inputs, viz:

$$\hat{x}_p = c_o + \sum_{j \neq p} c_j x_j \tag{4}$$

with c denoting a regression coefficient. The *RCC* between the residuals arising out of the Eq. (3) and Eq. (4) will now be free from the effects of input-input correlations, and is defined as the *PRCC* (Draper and Smith, 1981):

$$PRCC[y, x_p] = RCC[y - \hat{y}, x - \hat{x}_p] \tag{5}$$

We now consider a practical strategy for determining *PRCC*s that does not require building a sequence of regression models as suggested by Eq. (3)–(5). Following Iman et al. (1985), we write the augmented correlation matrix between the output variable, y, and the independent variables x_j, $j = 1,\ldots, n$, as:

$$\mathbf{C} = \begin{bmatrix} 1 & r_{12} & .... & r_{1n} & r_{1y} \\ r_{21} & 1 & .... & r_{2n} & r_{2y} \\ .... & .... & .... & .... & .... \\ r_{n1} & r_{n2} & .... & 1 & r_{ny} \\ r_{y1} & r_{y2} & .... & r_{yn} & 1 \end{bmatrix} = \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B^T} & \mathbf{1} \end{bmatrix} \tag{6}$$

where the matrix $\mathbf{A}$ represents the input-input correlation matrix with elements $r_{ij} = RCC[x_i,x_j]$, and the vector $\mathbf{B^T}$ denotes the output-input correlation vector with elements $r_{yj} = RCC[y,x_j]$. As shown by Rao (1973), the *PRCC* between x_j and y can be obtained from the elements of $\mathbf{C}^{-1}$, the inverse of $\mathbf{C}$, as follows:

$$PRCC[y, x_j] = -\frac{c_{jy}}{\sqrt{c_{jj} c_{yy}}} \tag{7}$$

where the subscript y is now used as the designator for row and column n+1 in $\mathbf{C}^{-1}$. It can also be shown that the *PRCC* and *RCC* are related as follows (RamaRao et al., 1998):

$$PRCC^2[y, x_j] = \left[1 - \frac{1}{1 + \left\{ RCC^2[y, x_j]/(1 - R^2) \right\}} \right] \tag{8}$$

where j is an index for the uncertain variable of interest, and R^2 denotes the coefficient of determination for the linear regression model with the j-th input variable included. Note that the importance ranking via *PRCC* and *RCC* will be identical for the case of uncorrelated inputs.

Helton (1993) describes several applications of the *PRCC* concept for analyzing the results of a probabilistic performance assessment model for the Waste Isolation Pilot Plant facility in Carlsbad, NM. RamaRao et al. (1998) showed that the square of the *PRCC* can be interpreted as the gain in R^2 of an input-output regression model – when the selected variable is brought into regression – as a fraction of the currently unexplained variance. They also presented results of a probabilistic sensitivity analysis using *PRCC*s for the proposed high-level radioactive waste repository at Yucca Mountain, NV.

3 ENVIRONMENTAL RISK ASSESSMENT MODEL

In what follows, the advantage of *PRCC*s for performing sensitivity analysis in a probabilistic model with correlated inputs is demonstrated using an analytical model of time-dependent risk arising from water-borne nuclide migration from a repository (Robinson and Hodgkinson, 1986). This simple "screening" model, which includes the most important aspects of radionuclide migration, contains components representing the source term, geosphere transport and biosphere transport for a single member radionuclide chain such as Technetium (Tc-99).

The source term is described by an initial containment time, T_o, followed by radionuclide release at a rate, k, proportional to the current inventory, with radioactive decay occurring all along. The time-dependent source flux, $S(t)$, after the containment period ($t > T_o$), is obtained as:

$$S(t) = kM_o e^{kt} e^{-(\lambda + k)t} \tag{9}$$

where M_o is the initial radionuclide inventory, and λ the radioactivity decay constant.

In order to deal with general inputs from the source term it is useful to calculate a Green's function for the geosphere, which gives the flux for a delta function input. For the one-dimensional transport case with advection, dispersion, equilibrium sorption and decay, the Green's function, $G(t)$, is given by:

$$G(t) = \frac{L e^{-\lambda t} e^{-R(L - vt/R)^2 / 4dvt}}{2(\pi dvt^3 / R)^{1/2}} \tag{10}$$

where d is the dispersivity, R the retardation factor, L the geosphere path length and v the groundwater velocity. The output flux from the geosphere, $F(t)$, is obtained via the convolution of $S(t)$ with $G(t)$, viz:

$$F(t) = \int_0^t S(t - \tau) G(\tau) \, d\tau \tag{11}$$

Finally, the biosphere path is assumed to be a stream which is the source of drinking water and hence the major exposure route for the critical group of human receptors. The biosphere conversion term, B, is simply a multiplication factor:

$$B = \frac{w}{W} q \zeta \tag{12}$$

where w is the annual amount of drinking water intake by an individual, W the stream flow rate, q the activity-to-dose factor, and ζ the risk factor for radiation induced cancer fatality.

The above equations can be combined using the Laplace transformation technique to yield a time-dependent consequence, $C(t)$, given by:

$$C(t) = \frac{1}{2} BkM_o \, e^{-\lambda t} \, e^{-Ld} \, e^{-RL^2/4dvt} \, e^{-vt/4dR}$$
$$\left[\phi\left\{ \left(\frac{RL^2}{4dvt} \right)^{1/2} + \left(\frac{vt}{4dR} - kt \right)^{1/2} \right\} + \phi\left\{ \left(\frac{RL^2}{4dvt} \right)^{1/2} - \left(\frac{vt}{4dR} - kt \right)^{1/2} \right\} \right] \tag{13}$$

where $\phi(z) = \exp(z^2)\mathrm{erfc}(z)$, and the other symbols are as defined previously. Note also that the consequence, $C(t)$, is essentially a risk term which expresses the probability of deaths per year - beyond the initial containment period ($t > T_o$).

4 EXAMPLE PROBLEM

The model described earlier is used to compute key uncertainty drivers of human health risk after 20,000 y of waste emplacement due to the migration of a single radionuclide from a hypothetical repository. The uncertain parameters in the model are: (1) fractional release rate, k, (2) groundwater velocity, v, and (3) biosphere conversion term, B. Each of these parameters is assigned a log-normal distribution with parameters as given in Table 1. Also tabulated therein are the fixed values assigned to all other parameters. Also, the correlation coefficient between $\log(k)$ and $\log(v)$ is specified as 0.50.

Table 1. Parameter distributions used in the example problem.

Parameter	Symbol	Distribution	Median Value	Std. Dev.
Initial inventory	M_o (Bq)	Fixed	5.0×10^{16}	--
Release rate	k (y^{-1})	**log-normal**	3.16×10^{-5}	0.333
Containment time	T_o (y)	Fixed	316	--
Decay constant	λ (y^{-1})	Fixed	3.25×10^{-6}	--
Retardation factor	R (-)	Fixed	10.0	--
Groundwater velocity	v (m y^{-1})	**log-normal**	0.1	0.167
Dispersivity	d (m)	Fixed	20.0	--
Geosphere path length	L (m)	Fixed	316	--
Biosphere conversion term	B (deaths Bq^{-1})	**log-normal**	1.0×10^{-18}	0.500

Note: Standard deviation is calculated for the $\log_{10}$-transformed parameters.

A Monte Carlo simulation was carried out using the above model and parameters, with 1000 Latin Hypercube samples utilized for uncertain propagation. The resulting cumulative distribution function (CDF) as calculated by Crystall Ball is shown in Fig. 1, exhibiting log-normal type characteristics with a P5-P95 range of ~10^{-3}-5 and a median value of ~0.2. Note that the outcome of interest is $C(t)$ as defined in Eq. (13), and normalized to a nominal value 10^{-6} deaths/y.

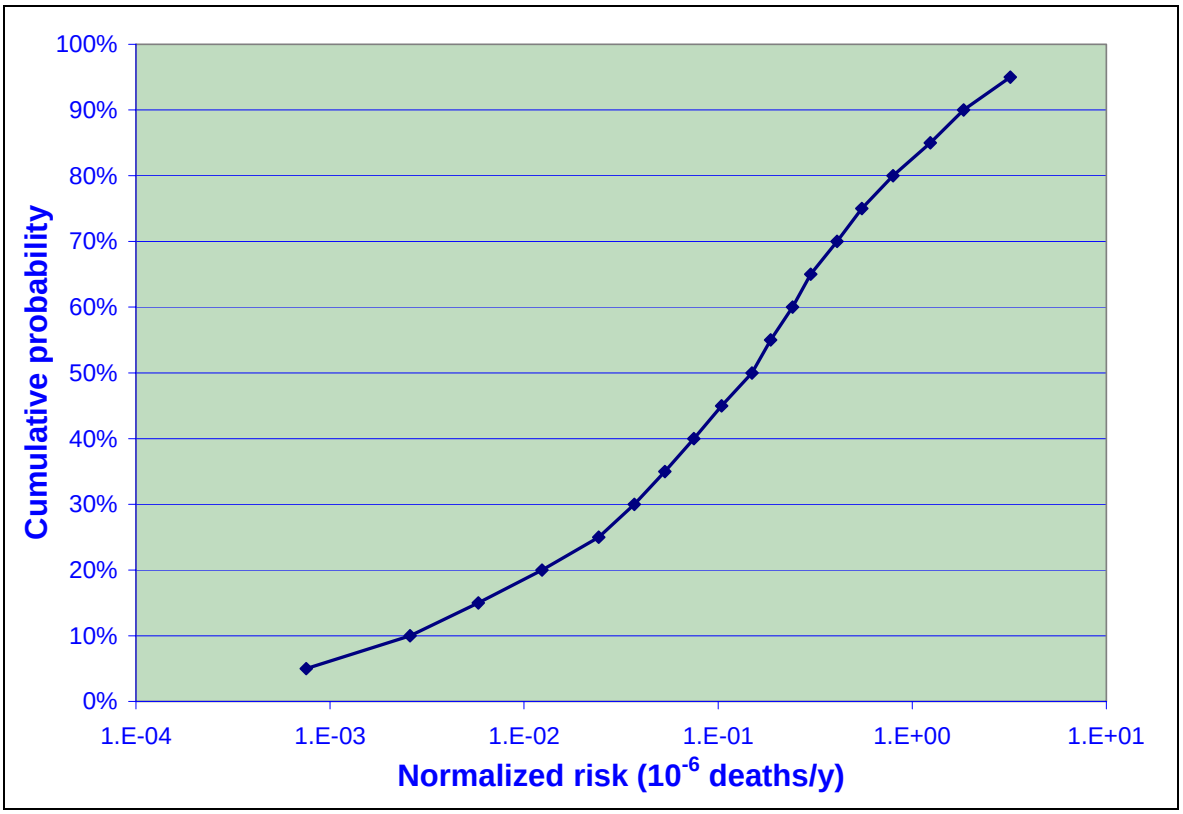

Figure 1: CDF of output

The sampled values of the inputs and the corresponding calculated value of the output for each of the 1000 realizations were extracted using Crystal Ball's Scenario Analysis utility. These values were then rank transformed, and used for calculating the input-input correlation matrix and the output-input correlation vector. The resulting augmented (rank) correlation matrix, with a structure similar to Eq. (6), is given below in Table 2.

Table 2. Augmented correlation matrix for example problem.

x1	X2	x3	y	
1	0.487004	-0.00525	0.60592	x1
0.487004	1	0.017091	0.82965	x2
-0.00525	0.017091	1	0.463146	x3
0.60592	0.82965	0.463146	1	y

Here, $x1$ denotes the fractional release rate, k, $x2$ denotes the groundwater velocity, v, $x3$ denotes the biosphere conversion term, B, and y denotes the normalized risk, $C(t)$. On the basis of the *RCC*s between the input and the output, the most important (sensitive) variable can be identified as $x2$ (v), followed by $x1$ (k) and $x3$ (B). These rankings are shown in Fig. 2 using a format similar to that of the sensitivity chart produced by Crystal Ball. It should be pointed out that the top two variables are correlated with a rank correlation coefficient of ~0.5.

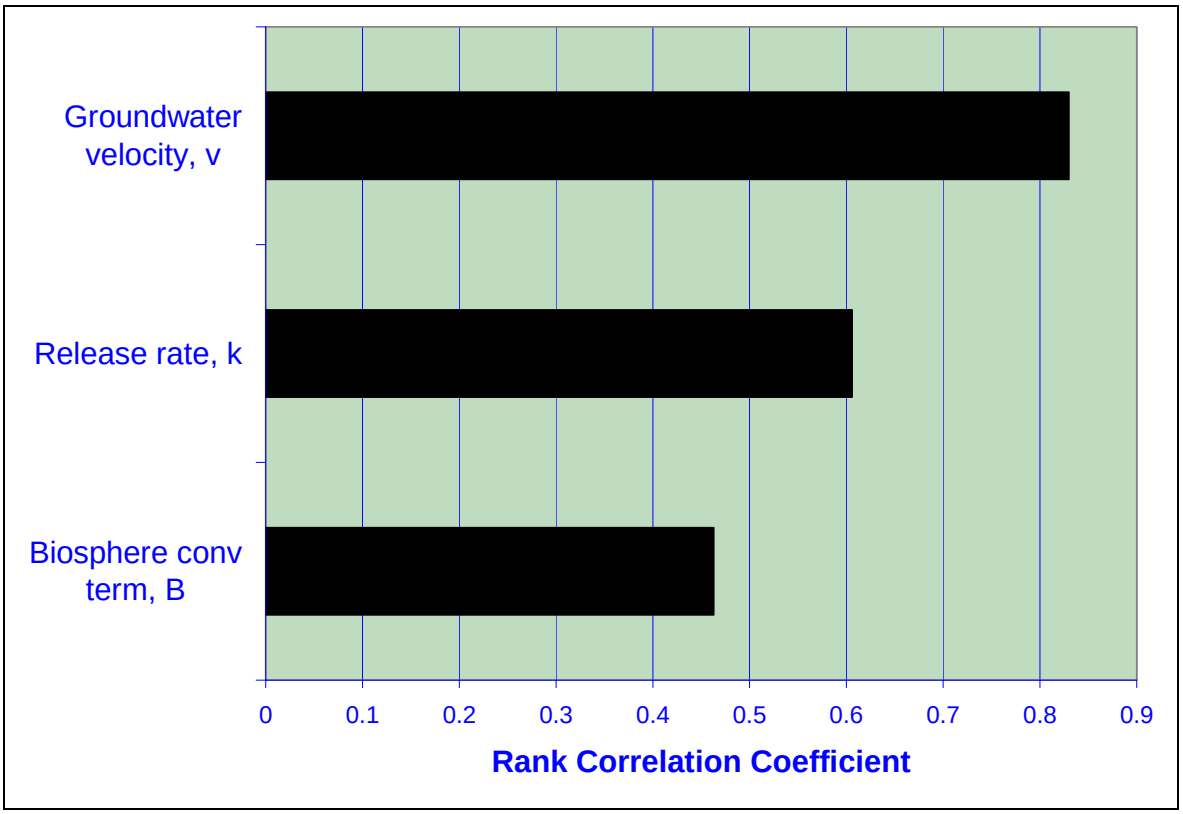

Figure 2: Sensitivity chart based on *RCC*

In order to determine the importance ranking using *PRCC*s, we first calculate the inverse of the augmented correlation matrix given in Table 2 using the Microsoft® Excel array function, MINVERSE, as follows:

Table 3: Inverse of augmented correlation matrix in Table 2.

2.705427	2.884248	2.332777	-5.1126
2.884248	10.21448	5.824164	-12.9195
2.332777	5.824164	4.844413	-8.48916
-5.1126	-12.9195	-8.48916	18.74822

The calculation of *PRCC*s is then carried out using the relationship given in Eq. (7). For example, the *PRCC* between $x2$ and y is calculated as:

$$PRCC[y,x_2]=-\frac{c_{2y}}{\sqrt{c_{22}c_{yy}}}=-\frac{-12.9195}{\sqrt{(10.21448)(18.74882)}}=0.934 \tag{14}$$

Similarly, $PRCC[y,x1]$ and $PRCC[y,x3]$ are calculated as .718 and .891, respectively. This suggests that the most important variable on the basis of PRCC is $x2$ (v), followed by $x3$ (B) and $x1$ (k). The corresponding sensitivity chart is shown in Fig. 3.

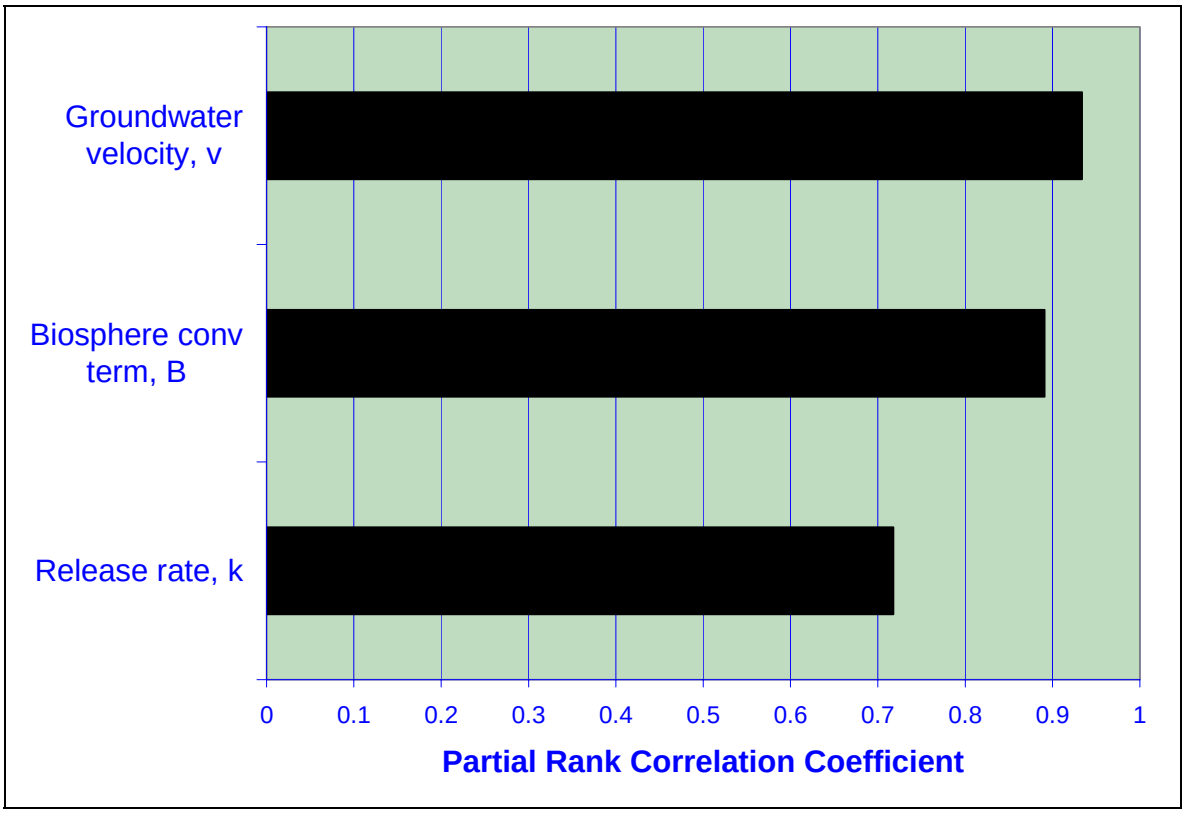

Figure 3: Sensitivity ranking based on *PRCC*

A comparison of the rankings based on *RCC*s and *PRCC*s shows that in both cases the most important input variable is groundwater velocity. However, the second most important variable suggested by *RCC*s, the fractional release rate, has a relatively high correlation to groundwater velocity. When this relationship is taken into consideration via the *PRCC*s, the true importance of the Biosphere conversion term is identified and it becomes the second most important variable.

The actual values of the *PRCC*s are not as easy to interpret as the *RCC*s, which are related to the slope of the best-fit line through a rank-transformed input-output scatter plot. While the relative magnitude of the *PRCC*s are important indicators of variable importance, the numeric values only have a specific meaning in the context of building a multivariate input-output regression model. As noted earlier, the square of the *PRCC* gives the increase in R^2, when a new variable is added, as a fraction of the currently unexplained variance in the model. From a practical standpoint, ranking the variables with *PRCC*s and examining scatter plots to understand input-output relationships would be a reasonable strategy for sensitivity analysis of probabilistic models when inputs are correlated.

5 CONCLUSION

This paper has presented a practical method for calculating sensitivity coefficients and uncertainty importance rankings for correlated inputs. The use of the partial correlation concept is well known in the linear regression and nuclear waste disposal safety analysis literature. Based on those sources, the paper describes how *PRCC*s can be computed readily using simple matrix algebra once the input-input correlation matrix and the input-output correlation vectors are obtained from the sampled values. It is hoped that the Crystal Ball users' community will find this methodology useful for identifying key drivers of uncertainty in spreadsheet-based probabilistic models where two or more uncertain inputs are correlated.

REFERENCES

Draper, N.R. and H. Smith, 1981, *Applied Regression Analysis*, 2nd Ed., John Wiley, New York, NY.

Helton, J.C., 1993, "Uncertainty and sensitivity analysis techniques for use in performance assessment for radioactive waste disposal," *Reliability Eng. & System Safety*, 42, 327-367.

Helton, J.C., J.W. Garner, R.D. McCurley and D.K. Rudeen, 1991, *Sensitivity Analysis Techniques and Results for Performance Assessment at the Waste Isolation Pilot Plant*, Sandia National Laboratories, Report SAND90-7103.

Iman, R.L., M.J. Shortencarier and J.D. Johnson, 1985, *A Fortran Program and User's Guide for the Calculation of Partial Correlation and Standardized Regression Coefficients*, Sandia National Laboratories, Report NUREG/CR-4122 / SAND85-0044.

RamaRao, B.S., S. Mishra, S.D. Sevougian and R.W. Andrews, 1998, "Uncertainty importance of correlated variables in a probabilistic performance assessment", Proc., *SAMO'98, Second International Symposium on Sensitivity Analysis for Model Output*, Venice, Italy, April 19-22.

Rao, C.R., 1973, *Linear Statistical Inference and its Applications*, 2nd Ed., John Wiley, New York, NY.

Robinson, P.C. and D.R. Hodgkinson, 1986, *Exact Solutions for Radionuclide Transport in the Presence of Uncertainty*, UK Atomic Energy Agency Report No. AERE R 12125.

Saltelli, A., K. Chan and M. Scott (editors), 2000, *Sensitivity Analysis*, John Wiley, London.

AUTHOR BIOGRPAHY

Dr. Srikanta Mishra (smishra@intera.com) is a Senior Engineer and Project Manager for Intera Inc., an environmental consulting company based in Austin, TX. He supervises the development and application of uncertainty analysis, statistical modeling, and risk assessment techniques for a wide variety of problems such as water resource management, nuclear waste disposal safety analysis, oil production forecasting and environmental health risk evaluation. Dr. Mishra previously worked as an academic researcher at Virginia Tech, as a manager of modeling and risk analysis projects for the Swiss and the U.S. radioactive waste disposal programs, and as an Adjunct Professor of Petroleum and Geosystems Engineering at University of Texas. He holds a PhD degree from Stanford University, an MS degree from the University of Texas at Austin, and a BTech degree from Indian School of Mines – all in Petroleum Engineering. Dr. Mishra is the author of over 100 publications including peer-reviewed journal articles, papers in edited conference proceedings and technical reports. He is the editor of the book "*Groundwater Quality Modeling and Management under Uncertainty*" published by ASCE, the American Society of Civil Engineers, and also the current Chair of the ASCE Ground Water Quality technical committee.