# Are the Regression Models Published in the Field of Environmental Science are Reproducible?

Tiaria Porche and Murty S. Kambhampati (Southern University at New Orleans, New Orleans, LA 70126),
Fred Rispoli and Vishal Shah (Dowling College, Oakdale, NY 11769), Timothy Green (Brookhaven National Laboratory, Upton, NY 11973)

## Abstract

This study focuses on the use of statistical analysis in research articles concerning environmental science which were published in various scholastic journals during 2004-2010. The main objective of this study was to validate the regression models presented in the published articles. A total of 266 biological articles were selected and analyzed. Screening for articles using regression analysis as the statistical method narrowed down the search to 26 articles. The data given in the 26 articles was used to develop the regression model using Microsoft Excel and compared to the published models. Only four of the articles have been validated 100% with the simulated/constructed models. On the contrary, the other twenty-two regression models failed to be reproduced using the published data. We found that in many publications the amount of information provided was not sufficient to reconstruct the model. In some publications, we believe that there were errors associated with the model development leading to non-reproducible models. The errors expressed may have occurred during the input of data into the software. Further studies are being carried out to understand the errors involved and elucidate the minimum information required to obtain the model from the publications. The advantage of being able to reproduce the given information validates if the statistical methods are used in the most effective and correct way possible.

## Introduction

- Many statistical methods have been used in environmental science journal publications either in concluding results or running an experiment.
- The common question that has began to arise in science is (1) how accurate are the results and interpretations and (2) from the obtained data could the experiment be reproduced?
- The present study focused on the validation of statistical methods formulated in recent publications in the field of environmental science.
- The use of multiple and linear regression analysis have been selected to validate the statistical methods used in the pervious publications.
- We considered the following parameters to evaluate the published regression models in each of 26 manuscripts in an attempt to reproduce the models: *x* and *y* variables, coefficients, p values, $R^2$ values, intercept, *n* (number of experimental runs), and experimental designs.
- To further investigate the significance of the data obtained in published models a Data Analysis tool was used in Microsoft Excel.

## Materials and Methods

- A total of 266 published articles were selected from various journals of environmental science.
- The Data Analysis tool was used to detect and validate the published models vs. the simulated models.
- We attempted to reproduce the data screened in the 26 articles, using the published models and data given in the articles.
- A linear and quadratic regression models were conducted.
- The y-values or a linear regression models are explained throughout each article. This type of regression analysis provided the information to form the conclusion if the relationship between the x and y values express a linear relationship. The methodology below will express a detailed understanding of the consistency between the two models.

*Regression Analysis*

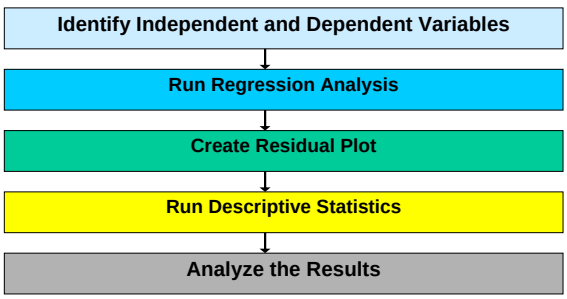
The following prediction equations were used to obtain the regression models:

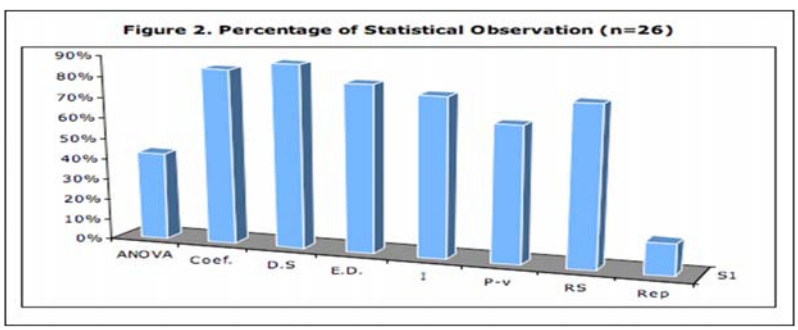
$Y = a + bx$ (linear)

$Y = a + bx + cx^2$ (multiple)

$Y = a + bx + cx^2 + dx^3$ (quadratic)

### Figure 1. Road Map Used to Simulate Published Regression Models




## Results, Discussion, and Conclusion

- 84.62% of the models failed and only 15.38% models were confirmed and reproducible.
- Further studies are being carried out to detect the specific errors of replicated model in comparison to the published models; identifying whether the errors occurred during input of data or the difference in the programs used.
- One of the most frequent issues that led to cause insufficiencies of manuscript rejection is a failure of authors to clearly and completely describes their statistical model.

| ID | Year | Journal | ANOVA | DS | ED | Rep | PM | SM |
|---|---|---|---|---|---|---|---|---|
| 1 | 2008 | Haz. Mat. | YES | YES | YES | NO | | |
| 2 | 2008 | Chem Eng. | | | | NO | | |
| 3 | 2009 | Haz. Mat. | YES | YES | YES | YES | 38.99 | 38.99 |
| | | | | | | | 0.99 9 | 0.98 |
| | | | | | | | -33.62 | -33.68 |
| | | | | | | | 0.99 | 0.99 |
| 4 | 2010 | Desalination | YES | YES | YES | YES | -172.38 | -172.28 |
| | | | | | | | 0.9726 | 0.9726 |
| 5 | 2006 | Haz. Mat. | NO | YES | YES | NO | | |
| 6 | 2010 | Chem. Eng. | NO | YES | YES | NO | | |
| 7 | 2007 | Chem. Eng. | NO | YES | YES | NO | | |
| 8 | 2010 | Chem. Eng. | YES | YES | YES | YES | 9.3167 | 9.3167 |
| | | | | | | | 0.9965 | 0.9965 |
| 9 | 2008 | Tetrahedron | YES | YES | YES | NO | | |
| 10 | 2010 | Sci. Total Env. | NO | YES | YES | NO | | |
| 11 | 2009 | Ocean Eng. | NO | YES | YES | NO | | |
| 12 | 2010 | Arabian Chem. | NO | YES | YES | NO | | |
| 13 | 2010 | Haz. Mat. | NO | YES | YES | NO | | |
| 14 | 2010 | Vaccine | NO | YES | YES | NO | | |
| 15 | 2007 | Haz. Mat. | YES | YES | NO | NO | | |
| 16 | 2010 | Haz. Mat. | NO | NO | NO | NO | | |
| 17 | 2010 | Atmos. Envi. | YES | NO | YES | NO | | |
| 18 | 2010 | Biore. Tech. | NO | YES | NO | NO | | |
| 19 | 2007 | Ecotox. Env. Safety | YES | YES | YES | NO | | |
| 20 | 2008 | Analyt. Chimi. | NO | YES | NO | NO | | |
| 21 | 2009 | Chem. Eng | YES | YES | YES | NO | | |
| | | | | | | | 55.7 | 55.7 |
| | | | | | | | 0.97 | 0.97 |
| 22 | 2010 | Haz. Mat. | NO | YES | YES | YES | 51.6 | 51.6 |
| | | | | | | | 0.98 | 0.98 |
| | | | | | | | 50.9 | 50.9 |
| | | | | | | | 0.92 | 0.92 |
| | | | | | | | 81.2 | 81.2 |
| | | | | | | | 0.98 | 0.98 |
| 23 | 2004 | Ecotox. Env. Safety | YES | NO | YES | NO | | |
| 24 | 2010 | Haz.Mat. | NO | YES | YES | NO | | |
| 25 | 2010 | Biotech. Prog. | NO | YES | NO | NO | | |
| 26 | 2008 | Biotech. Prog. | YES | YES | YES | NO | | |

DS: Design Software; ED: Experimental Design; Rep.: Reproducibility; PM: Publication Model; SM: Simulation Model

## Necessary Information In Assessing Statistical Simulations

- Specific field of study
- Design of experiment used, if any (i.e. Mixture Design, Box-Behnken, Plackett Burman)
- Consideration of p-values
- Design Software use in construction of statistical models
- The number of replications used in statistical analysis
- Identification of statistical errors, if any (mean, median, error of measurements, percent error).

## Limitations in Reproducibility of Regression Models

- Experimental Errors — Errors of Measurements
- Insufficient Data — p-values >0.05
- Inadequate software design used — Unidentified number of replications

## References

[1] Annadurai G. et. al 2008. Haz. Mat. 151.171
[2]Anuniata O. A. and Cussa J. 2008 Chem. Eng. 138. 510
[3]Ay F. et. al. 2009. Haz. Mat. 160.230
[4] Baskan. M. B. and Pala A. 2010. Desalination. 254. 42
[5] Brasil J.L. et. al. 2006. Hat. Mat. B133.143
[6] Chuan. L. et.al. 2010. Chem. Eng. 165.482
[7] Djoudi. W. et. al. 2007. Chem Eng. 133. 1
[8] Dopar M et. al. 2010. Chem. Eng.CEJ 7327. 1
[9] Glasnov. T. N et. al. 2008.Tetrahedron. 64. 2035
[10] Homem. V. et. al. 2010. Sci. Total Enviorn. 408.6272
[11] Islam M.F. and Lye L.M. 2009. Ocean Eng. 36.237
[12 ]Jabri. M. et. al. 2010. Arabian Chem. (*in press*)
[13] Khajeh A. and Modarress H. 2010. Haz. Mat. 179.715
[14] Kutle L. et. al 2010. Vaccine. 28.5497
[15] Landaruru J. et. al. 2010. Haz. Mat.180.524
[16] Lima. E. C. et. al 2007. Haz. Mat.140.211
[17] Mabilia R. et.al 2010. Atmospheric Enviornment. 44.3942
[18] Mohajeri L. et.al 2010. Bioresource Technology.101.893
[19] Mohan S.V.et.al. 2007. Ecotoxicology and Enviorn. Saftey. 68.252
[20] Molina M. et.al. 2008. Analytica Chimica Acta. 628.155
[21] Prasad R.K. and Srivastava S.N. 2009. Chem. Eng. 146.22
[22] Sakkas. V.A. et.al.2010.Haz. Mat. 175.33
[23] Ren S. et. al. 2004. Ecotoxicology and Environ. Saftey. 59. 38
[24] Rispoli F. et.al. 2010. Haz. Mat. 180.212
[25] Rispoli F. et.al. 2010. Biotechnol.Prog.26.938
[26] Rispoli F. and Shah V. 2008.Biotechnol.Prog. 214.648
[27] Yim, K.H. et.al. 2010.Korean J Pain.23.35
[28] Strasak, A. M. et. al. 2007. Swiss Med Weekly. 137.44
[29] Moye', L.A. 2006. Statistical Reasoning in Medicine: The Intuitive P-value Primer. 275

## Acknowledgements

We thank NSF (grant # HRD-0928797 and DUE-0806894) and DOE/BNL for financial support and facilities. We also thank Noel Blackburn (FaST Program Manager), OEP staff and fellow student interns for their help and support at various stages of the project.