

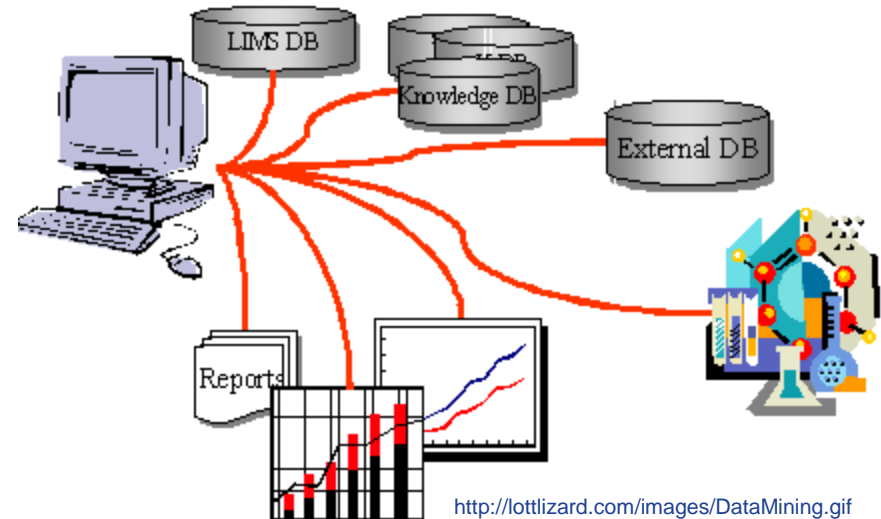
Data Mining and LCA: A Survey of possible marriages

Matthew Pietrzykowski, GE Global Research
pietrzyk@research.ge.com

Life Cycle Assessment IX, Computational Methods
Boston, October 2, 2009



www.liacs.nl/~edegraaf/img/datamining.jpg



<http://lottlizard.com/images/DataMining.gif>

http://www.science-environnement.ch/en/gfx/lifecycle_illu_en.jpg



imagination at work

Acknowledgements

GE Global Research Ecoassessment Center of Excellence

William Flanagan, Ron Wroczynski, Angela Fisher

Inspirational discussions with:

Professor Mark Huijbregts, Radboud University

Professor Sangwon Suh, University of Minnesota

Outline

Introduction

Data Integrity

1. Collection errors?
2. Data gaps or input issues?

Exploratory Data Analysis (EDA)

1. Multivariate Data Visualization
2. LCI natural structure prior to subjective aggregation

Sensitivity Analysis

1. Global vs. local methods
2. Typical stochastic approaches (Monte Carlo Analysis)
3. Alternatives? (DACE, Elementary Effects, Bayesian, etc..)

Summary

Life Cycle Interpretation – Methods and Opportunities

Analytic Methods

- **Matrix Perturbation Theory¹** – uncover non-linear sensitivities from a system of linear equations

Numerical Methods

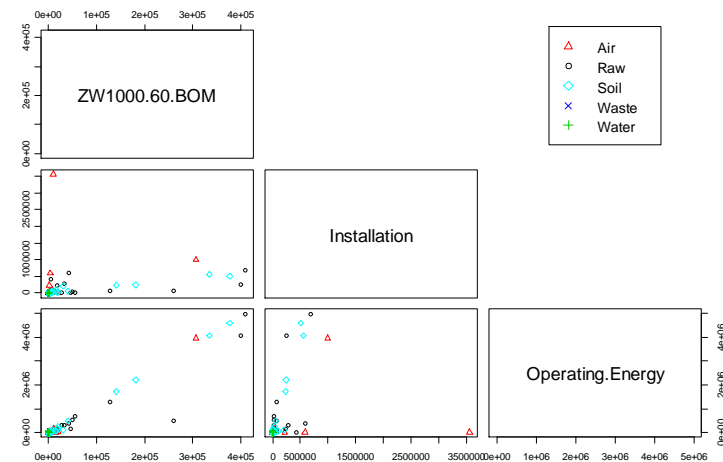
- **Monte Carlo** – N model executions with randomized inputs to uncover model uncertainty and sensitivity
- **DACE** – Design of Computer Experiments
- **Global Sensitivity Analysis**

Data Mining Opportunities

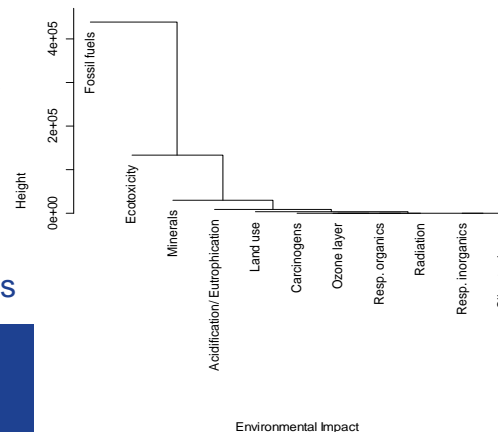
- Exploratory Analysis to understand the underlying structure of the data flows
- Outlier Detection
- Inventory/Process Contribution/Impact Structural Analysis
- Some techniques to employ
 - o Principal Component Analysis / Factor Analysis
 - o Discriminant Analysis/Cluster Analysis
 - o Multivariate Regression (MLR, PCR, PLS, Ridge Regression)
 - o Neural networks, Bayesian approaches

LCA literature is sparse on this topic; perhaps an opportunity to contribute?

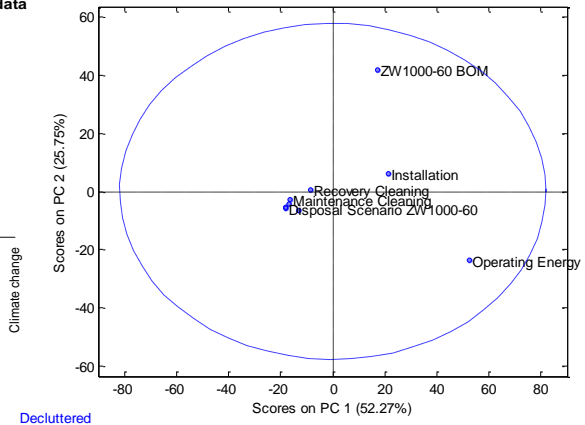
Matrix plot exploring significant Unit Process relationships



Hierarchical Clustering based on impact assessment data



Samples/Scores Plot of Zenon Cartridge Facility Life Inventory.txt



¹Heijungs, R. Suh, S., (2002) *The Computational Structure of Life Cycle Assessment*, Kluwer Academic Publishers: Dordrecht, The Netherlands

Data Example

LCA Energy Grid Mixes by Geographic Region

- Goal and scope required examination of the effects of the electricity grid mixes on the operating energy by geographic region
- Additional unit processes were developed to reflect grid mixes in geographic regions of interest
- Once source data was located and input in the system, the data integrity was checked

Data Integrity

- Not to be confused with Data Quality by LCA definition
- LCA data collection and input issues
 1. Human error
 2. Data gaps (flaws in the collection process)
- LCA data is heterogeneous:
 1. Multiple sources
 2. Different quality standards
- Data Mining offers tools to identify and potentially correct data integrity issues

Data Integrity Example

Identify inconsistent data

Troubleshooting/Debugging

Software Required:

Statistical packages like:
Matlab, R, Minitab, Excel with add-ins, etc

Techniques Used:

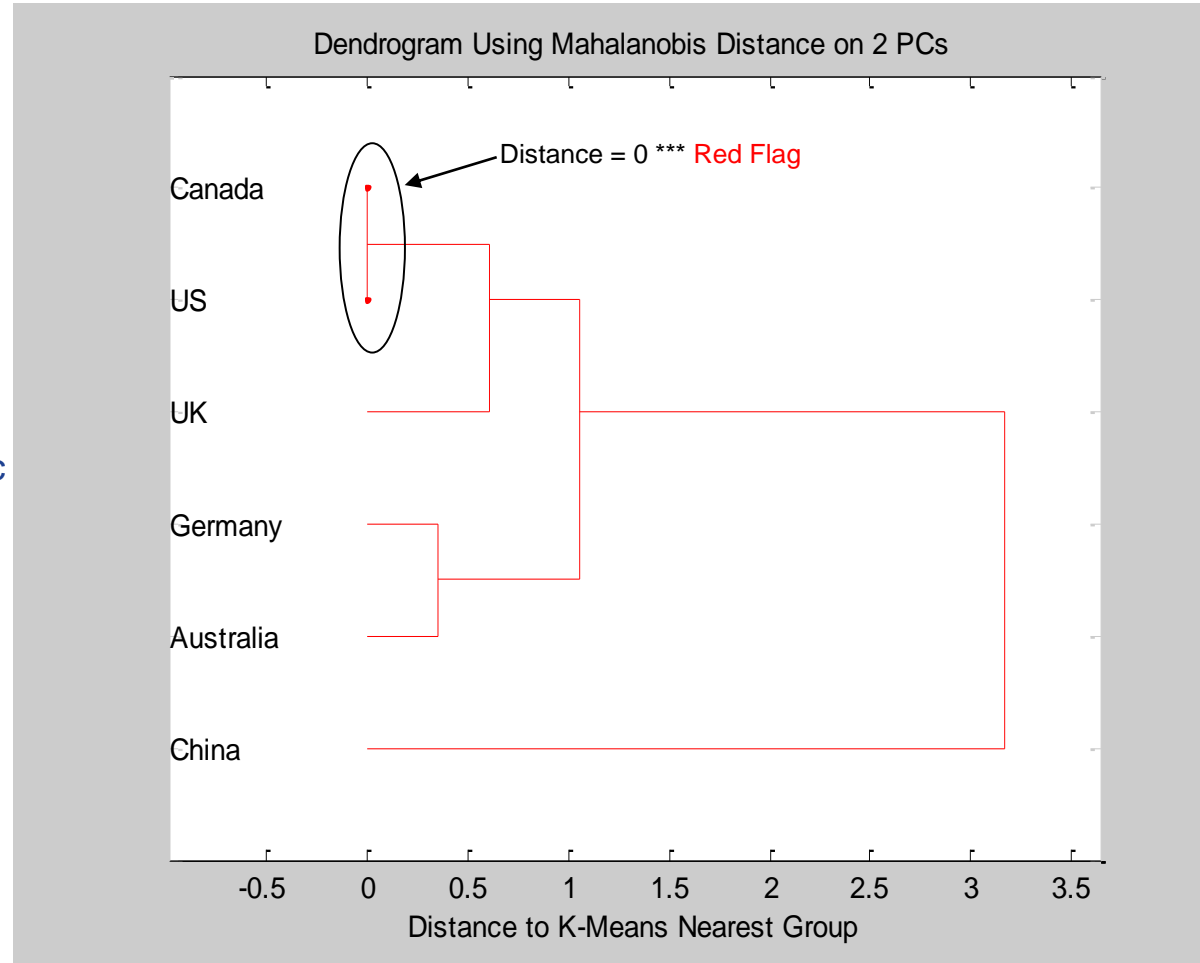
1. K-Means Clustering

Compare group means by
some distance measure

Large Distance = Different
Small Distance = Same

2. Dendrogram

Visualize groups relative to
their distances



Exploratory Data Analysis (EDA)

Multivariate Data Visualization

Adhere to the golden rule of data analysis, **LOOK AT YOUR DATA**

Uncover inherent structure due to

Natural variation

- Underlying covariances and correlations
- Does the data naturally favor a few outlying observations?

Natural grouping

- Common categorical or ordinal variables across the data

Ask questions like:

Is the variation in my data skewed? If so, why? How might this propagate as error?

Are there local pockets (groups) that correlate with more variation than others?

EDA LCA Example: Energy Grid Mixes

- Raw inventory data is skewed
- Process data spreads 6 decades!
- Substance data spreads 11 decades!

Transform the data for scale and symmetry

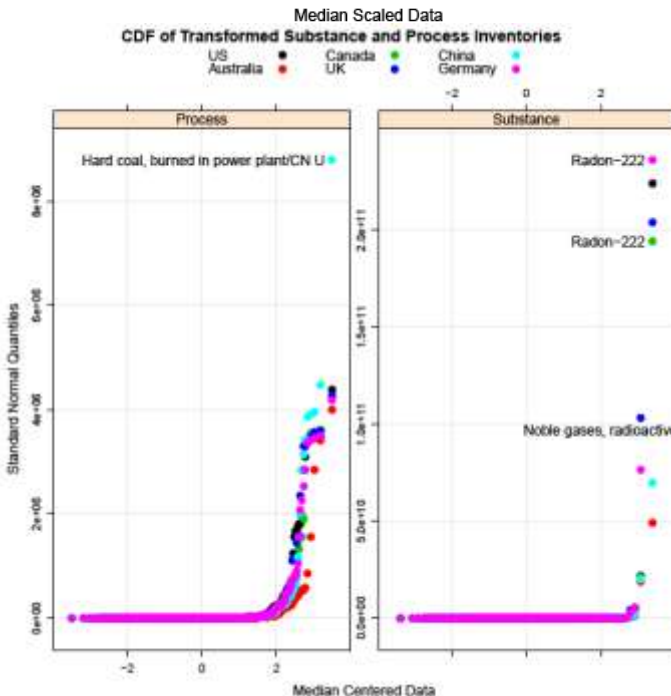
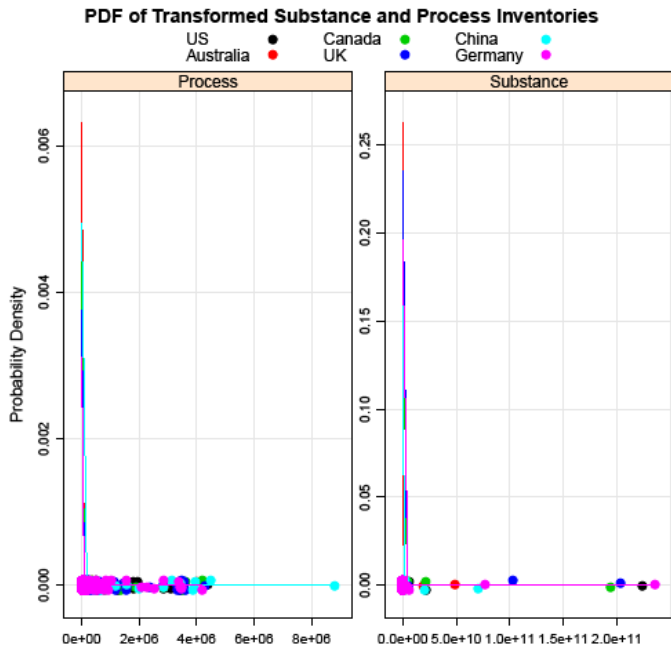
Typical Transformations:
 Log(X)*, Sqrt(X)*, 1/X**

*data must be positive so shift data by adding a scaler

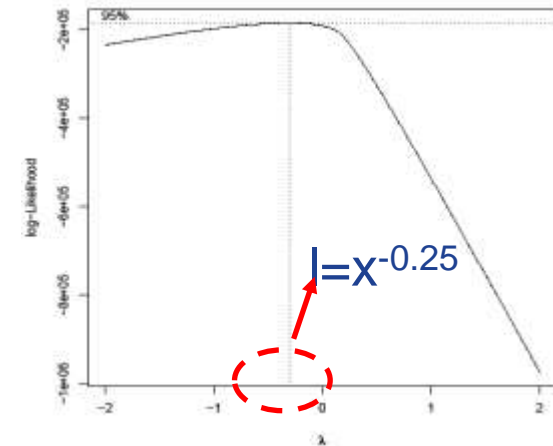
**data will be “flipped”, so it is necessary to reflect by multiplying by -1

Box-Cox Family of transformations:

$$y_i^{(\lambda)} = \begin{cases} \frac{y_i^\lambda - 1}{\lambda(GM(y))^{\lambda-1}}, & \text{if } \lambda \neq 0 \\ GM(y) \log y_i, & \text{if } \lambda = 0 \end{cases}$$



Boxcox Transform Results



EDA LCA Example: Energy Grid Mixes

Transformed data preserves the order of the original data

** When transforming data, “shift” the data to all positive values by a scalar quantity. PRESERVE ORIGINAL ORDER

Software Required:

Statistical packages like:

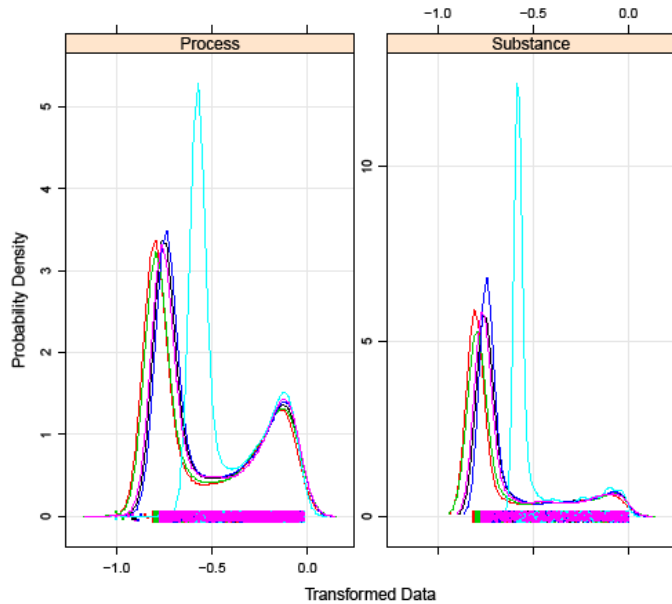
Matlab, R, Minitab, Excel with add-ins, etc

Techniques:

- 1) Lattice Graphics
- 2) Cumulative Distribution plots
- 3) Probability Density plot

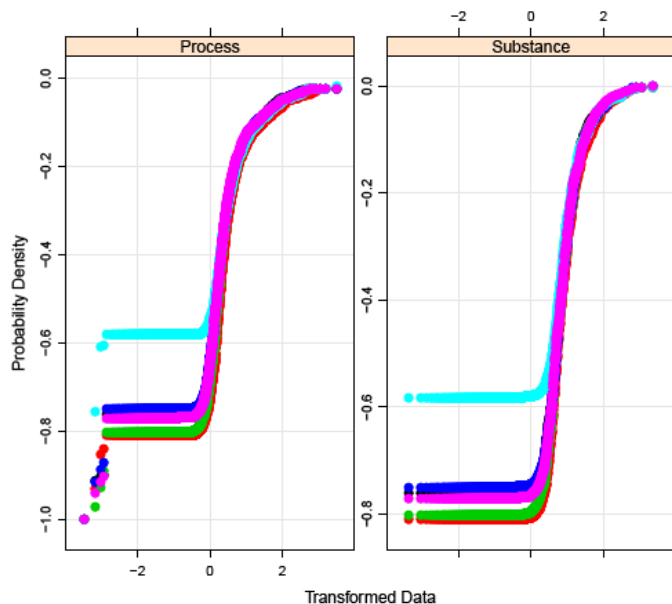
PDF of Transformed Substance and Process Inventories

US ● Canada ● China ●
Australia ● UK ● Germany ●



CDF of Transformed Substance and Process Inventories

US ● Canada ● China ●
Australia ● UK ● Germany ●



Transformations are useful because:

Data is more symmetric and the scale is manageable

Are there outliers?

Are there any group differences emerging?

EDA: Multivariate Data Visualization

Data:

Transformed inventory data

Software Required:

Statistical packages like:

Matlab, R, Minitab, Excel with add-ins

Techniques:

- 1) Matrix Plots
- 2) Hierarchical Clustering (Unsupervised Classification)
 - agglomerative or divisive
 - defined by its linkage



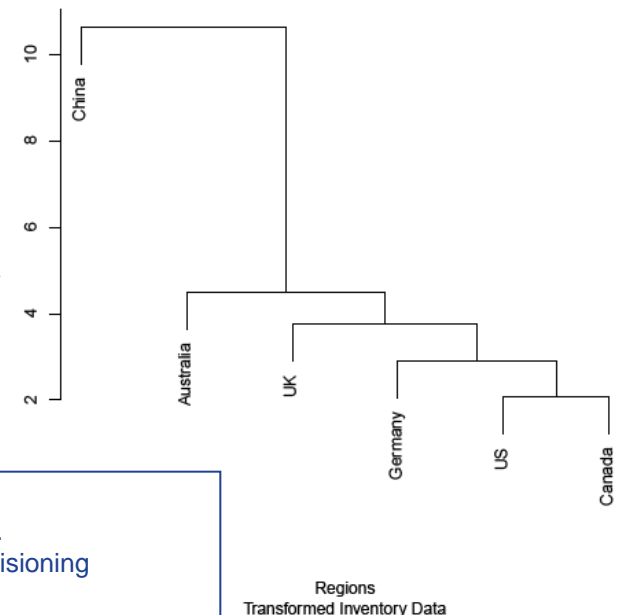
Gross Relationships ?

What about aggregation?

How about outliers?

How will the geographic correlations affect my Uncertainty?

Unsupervised Hierarchical Clustering



Unsupervised (no help from predefined groups) therefore *Independent* of the impact assessment method

Does one region stand out compared to the rest?

Is difference between China and the rest a modeling issue or true differences in the data?

Is there a way to plot both group and inventory information?



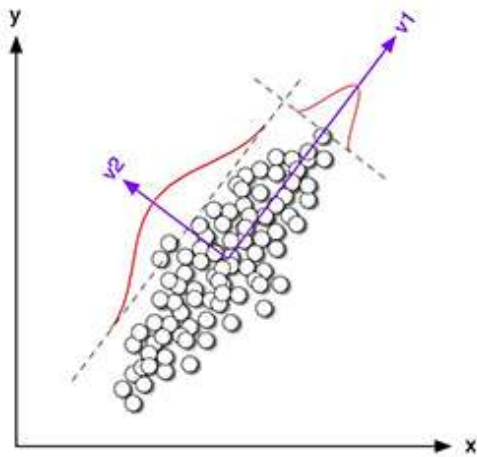
See:

Cleveland, W.S., *Visualizing Data* (Summit, NJ. 1993, Hobart Press.

Tufte, E.R., *The visual display of quantitative information*. 1983. Envisioning information, 1990.

EDA: Inherent Structure using Principal Component Analysis (PCA)

PCA



http://www.cs.cornell.edu/courses/cs322/2008sp/images/thumb_PCA.png

Techniques Used: Principal Component Analysis Biplots

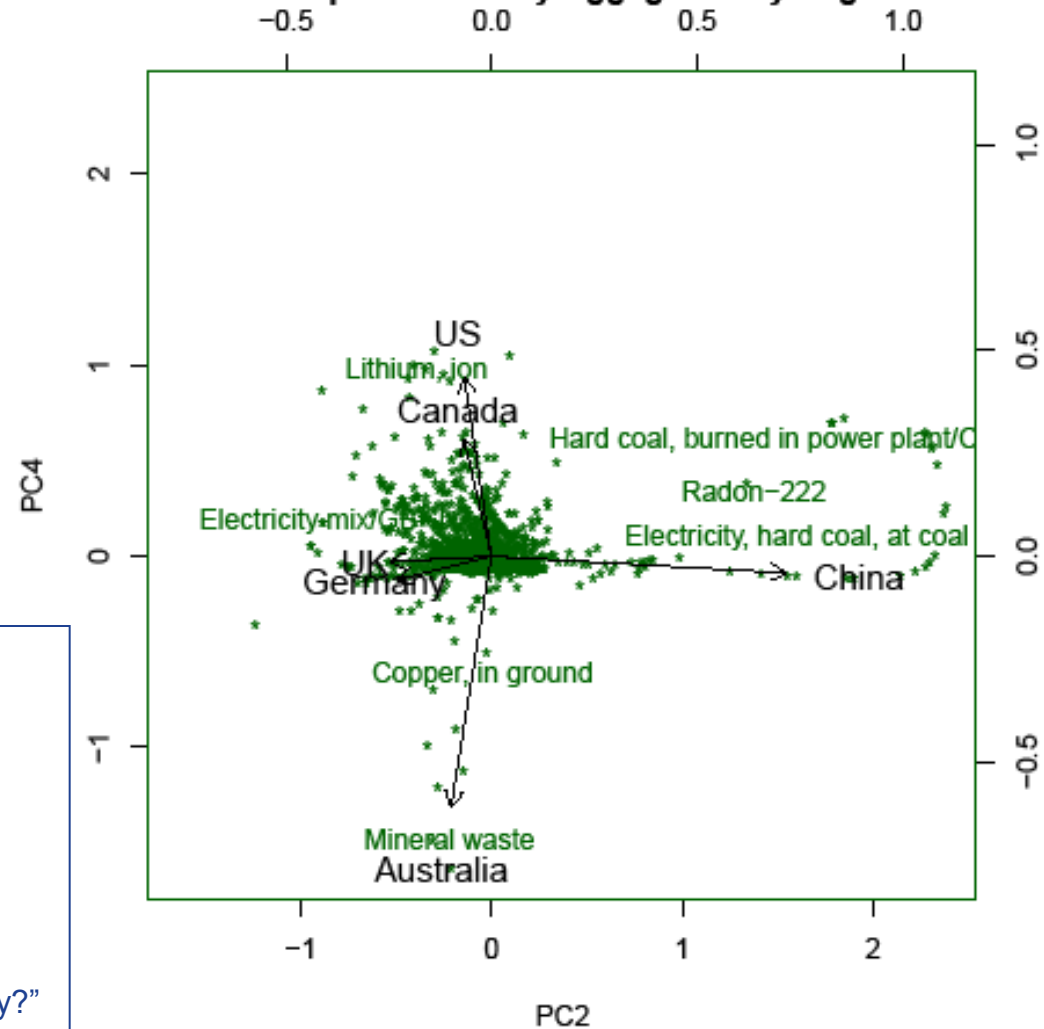
Variance based data reduction technique

Results in:
Orthogonal (uncorrelated) latent variables

Reduced dimensionality – 2D plots can be used to visualize patterns

Link groups to underlying data (inventory) thus answers, “What is so different with China’s inventory?”

PCA Biplot: Inventory Aggregation by Region

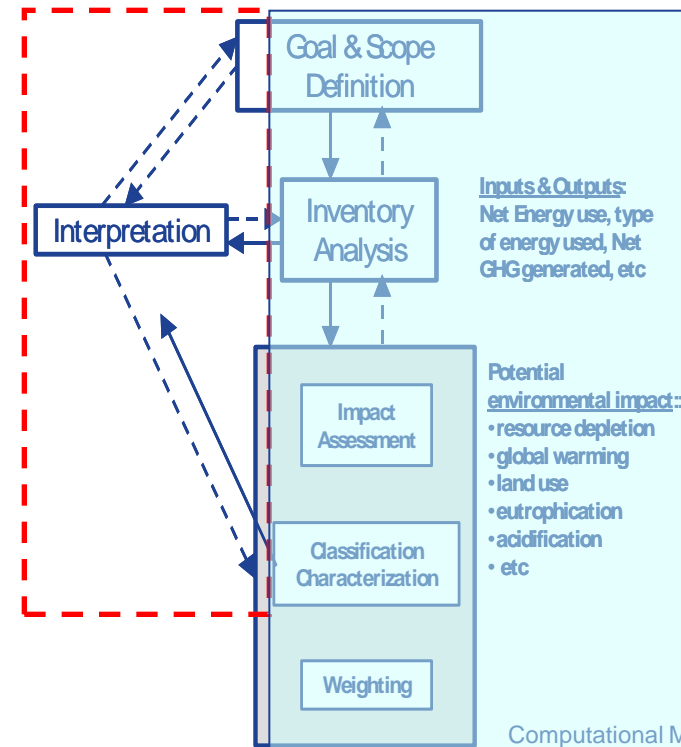


EDA Summary

- Visualize your data... it's there and has a lot to say
- Understand the unique characteristics of your inventory ... are your observations the same after applying an impact assessment method?

Model building, structure and architecture are big steps ... so too are the analyses

Life Cycle Procedure



Sensitivity Analysis

Sensitivity Analysis is a common and frequent practice in LCA

How much do we really know about it and what options are out there?

Global (Sampling, Designs, etc) vs. Local (Derivatives)

More than just One-at-a-Time (OAT) and Monte Carlo methods

Benefits of global methods:

- **Identify significant inputs AND**
- **Assess interacting terms**
- **Assess nonlinearity**
- **More “bang for the buck”**
 - **meta-models (“play in the sandbox”)**
 - **optimization (Is my LCA as robust as it can be?)**

Sensitivity Analysis Example

DACE (Data Analysis of Computer Experiments)

- Different than standard DOE methods
- Focuses on sampling strategy relative to the questions being asked:
 - Response Surface
 - Latin Hypercube
 - Hammersly designs
- Allows structure, efficiency, investigation of collinear and higher order effects
- Can be used to build meta-models, screening, or to identify more robust models

Sensitivity Analysis Example

General Factorial Design: Geography, Facility Life and Product Life

From my impact assessment, Respiratory Inorganics is highly impacted.

Can I assess the sensitivity of Respiratory Inorganics on my three factors?

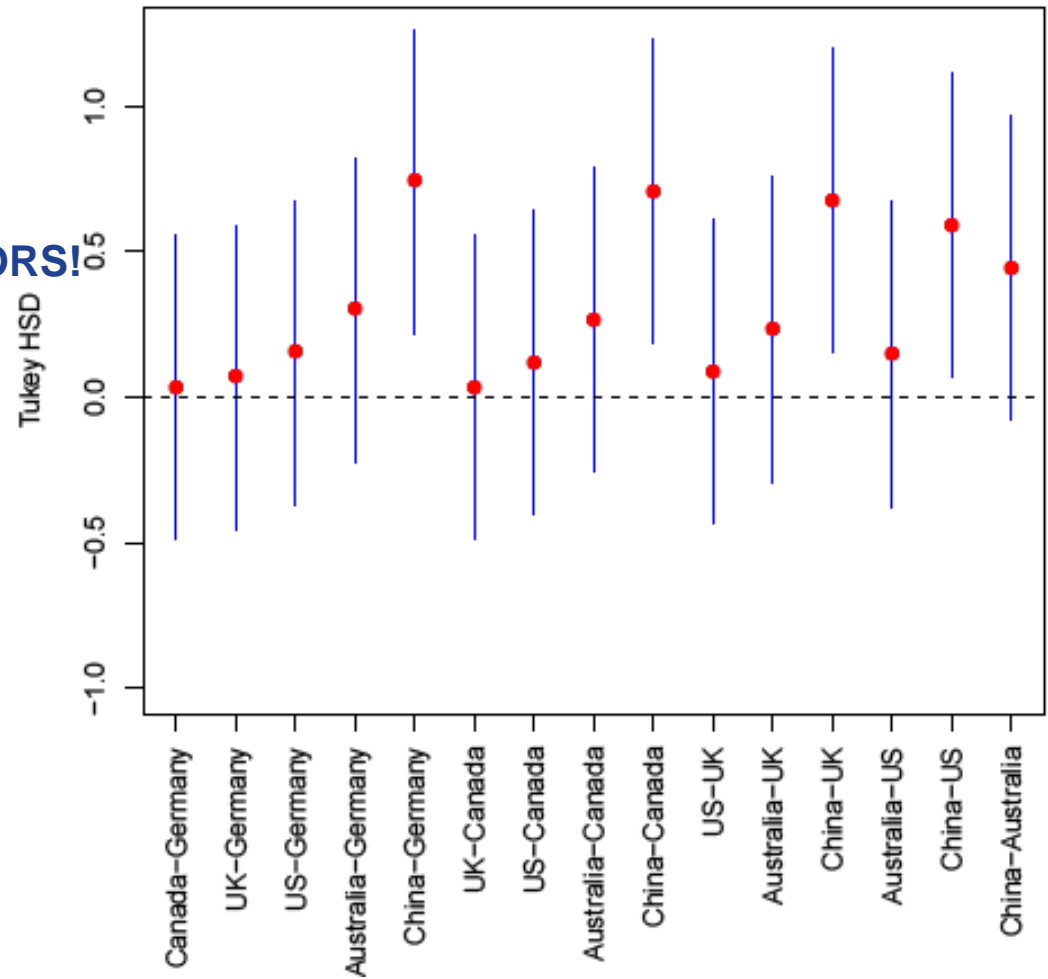
Assess significance
IN THE PRESENCE OF OTHER FACTORS!

Factor	Type	Levels	Values
Facility Life	fixed	2	10, 50
Product Life	fixed	2	2.5, 15.0
Geography	fixed	6	1, 2, 3, 4, 5, 6

Analysis of Variance for 1/sqrt(Resplnorg)

Source	DF	Seq SS	Adj SS	Adj MS	F	P
Facility Life	1	3.20348	3.20348	3.20348	305.53	0.000
Product Life	1	0.42554	0.42554	0.42554	40.58	0.000
Geography	5	1.32610	1.32610	0.26522	25.29	0.000
Error	16	0.16776	0.16776	0.01049		
Total	23	5.12288				

Significant Geographies by TukeyHSD



Sensitivity Analysis Example

Response Surface: Facility Life and Product Life

Meta-Modeling What if I want to explore China a bit more...can I build a meta-model for the design space?

Response Surface Regression: Resp. inorga versus Facility Lif, Product Life

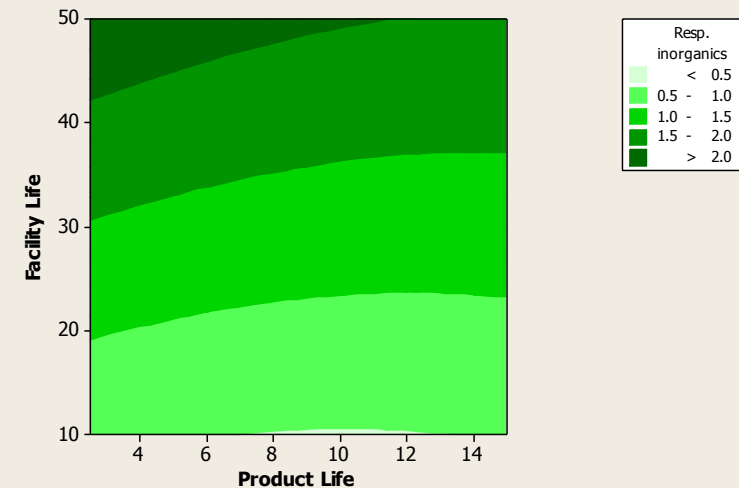
Term	Coef	SE Coef	T	P
Constant	0.245895	0.072547	3.389	0.043
Facility Life	0.044973	0.004116	10.927	0.002
Product Life	-0.036540	0.012506	-2.922	0.061
Facility Life*Facility Life	-0.000000	0.000064	-0.000	1.000
Product Life*Product Life	0.002088	0.000656	3.182	0.050
Facility Life*Product Life	-0.000609	0.000145	-4.200	0.025

S = 0.03625 R-Sq = 99.9% R-Sq(adj) = 99.7%

Analysis of Variance for Resp. inorganics

Source	DF	Seq SS	Adj SS	Adj MS	F	P
Regression	5	3.88667	3.886668	0.777334	591.54	0.000
Linear	2	3.85018	0.179163	0.089581	68.17	0.003
Square	2	0.01331	0.013305	0.006653	5.06	0.109
Interaction	1	0.02318	0.023180	0.023180	17.64	0.025
Residual Error	3	0.00394	0.003942	0.001314		
Total	8	3.89061				

Contour Plot of Resp. inorganics vs Facility Life, Product Life



Meta Model:

$$\text{Respiratory Inorganics} = \beta_0 + \beta_1 * \text{Facility Life} + \beta_2 * \text{Product Life} + \varepsilon$$

$$\text{Rsplnorg} = 0.3 + 0.04 * \text{Facility Life} - 0.02 * \text{Product Life} + \text{Error}$$

Sensitivity Analysis Example

Response Surface: Facility Life and Product Life

Meta-Modeling

What if I want to explore China a bit more...can I build a meta-model for the design space?

Response Surface Regression: Resp. inorga versus Facility Lif, Product Life

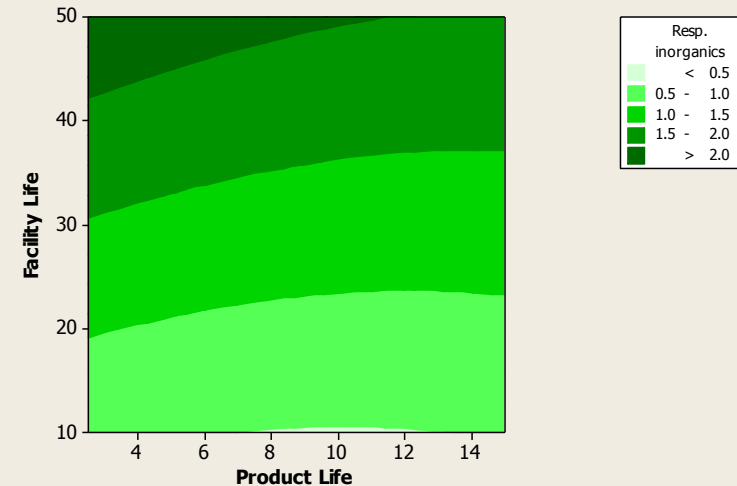
Term	Coef	SE Coef	T	P
Constant	0.245895	0.072547	3.389	0.009
Facility Life	0.044973	0.004116	10.927	0.000
Product Life	-0.036540	0.012506	-2.922	0.028
Facility Life*Facility Life	0.000000	0.000064	-0.000	1.000
Product Life*Product Life	0.002088	0.000656	3.182	0.021
Facility Life*Product Life	-0.000609	0.000145	-4.200	0.008

S = 0.03625 R-Sq = 99.9% R-Sq(adj) = 99.7%

Analysis of Variance for Resp. inorganics

Source	DF	Seq SS	Adj SS	Adj MS	F	P
Regression	5	3.88667	3.886668	0.777334	591.54	0.000
Linear	2	3.85018	0.179163	0.089581	68.17	0.000
Square	2	0.01331	0.013305	0.006653	5.06	0.021
Interaction	1	0.02318	0.023180	0.023180	17.64	0.008
Residual Error	3	0.00394	0.003942	0.001314		
Total	8	3.89061				

Contour Plot of Resp. inorganics vs Facility Life, Product Life



Meta Model:

$$\text{Respiratory Inorganics} = \beta_0 + \beta_1 * \text{Facility Life} + \beta_2 * \text{Product Life} + \varepsilon$$

$$\text{Rsplnorg} = 0.3 + 0.04 * \text{Facility Life} - 0.02 * \text{Product Life} + \text{Error}$$

Overall Summary

This discussion was meant to introduce concepts and tools that can aid the LCA analyst to produce the best models possible

The topics and examples covered are NOT exhaustive

Areas not discussed but ones that could be impacted are:

- Data scaling, uncertainty and Bayesian approaches
- Subjective decisioning (loss functions, dissimilarity measures)
- Qualitative Data Analysis (recursive abstraction, content analysis, analytic induction)
- Global Sensitivity Analysis (Elementary Effects, LHS sampling, Variance Measures)