Addressing Challenges for Knowledge Discovery from Data in the Domain of Seaport Integration

Ana Ximena Halabi Echeverry and Deborah Richards

Department of Computing
Macquarie University
NSW 2109, Australia
{ana.halabiecheverry,deborah.richards}@mq.edu.au

Abstract. Discovering knowledge from data for decision making is dependent on the existence of data relevant to the decision at hand. For decisions in domains that involve many different factors and concerns, such as seaport integration, data may exist across many repositories managed by different organizations with different goals and foci, not to mention different data structures, entities, labels, units of measurement, categories and time periods. To use this data for decision making, approaches to combine the data and handle missing values are two of the problems, among others, that need to be addressed. In this paper we discuss the need for managing micro and macro-level data and our approach to handle missing values.

Keywords: Seaport Integration, Data Aggregation, Missing Values.

1 Introduction

Discovering knowledge from data for decision making is dependent on the existence of data relevant to the decision at hand. For decisions in domains that involve many different factors and concerns, such as seaport integration, data may exist across many repositories managed by different organizations with different goals and foci, not to mention different data structures, entities, labels, units of measurement, categories and time periods. In this paper we present an approach to address two key issues which will affect the quality of decision making in seaport integration and other domains: data aggregation and missing values. We further discuss the notions of macro and micro data to allow strategic/high-level decision making to be conducted when only operational/low-level data is available. In Section 2 we discuss the need to aggregate data from multiple sources and the role of macro and micro data to support strategic and complex decision making. In Section 3 we consider how to handle missing values in the context of identification of ports who were leaders in compliance with environmental standards. Conclusions and future work appear in Section 4.

2 Aggregating Data from Multiple Sources

Port authorities (PAs) tend to be concerned with operational decisions and have tended to make local decisions [8, 9]. However, the increasingly competitive global

D. Richards and B.H. Kang (Eds.): PKAW 2012, LNAI 7457, pp. 73–85, 2012.

environment demands that PAs engage in longer-term and higher-level decision making be undertaken. Key reasons why strategic decision making does not occur includes the lack of available data and models or approaches to analyse the data. In our investigations concerning seaport integration, it became quickly apparent that potentially relevant data exists in many different locations. This data may use different labels/names, units of measurement and time frames. Some concepts may overlap [1] and be difficult to match. We see in Figure 1 examples of data from just four of the relevant sources in the US seaport domain: U.S. Army Corps of Engineers, U.S. Census Bureau, US Department of Homeland Security and US Department of Transportation. Each of those repositories offers a hierarchical structure or set of modules of information, which address a certain level of decision-making for each individual institution. If a seaport authority wishes to make any decision by analysing those data sets, the process will involve disaggregate analysis that unavoidably results in losing various degrees of information. In Figure 1 the data gathered/supplied by the US Census Bureau represents aggregated and summarised data (i.e. macro level data and abstract/high level concepts such as "people and households" and "geography"). Different colours indicate that some variables concern different types of decisions and different subsystems (discussed further below) that comprise the seaport domain.

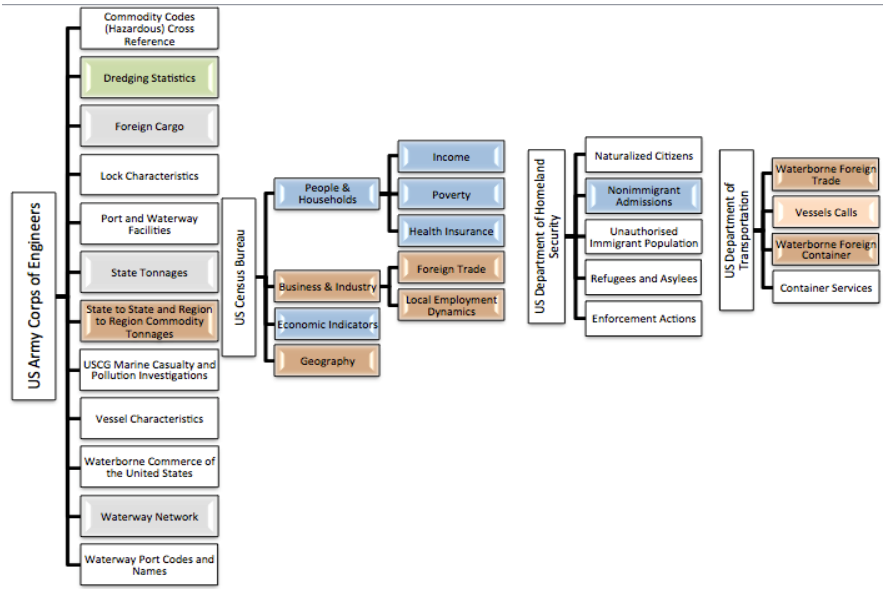

Fig. 1. Macro data repositories on the US Data websites[1]

[1]US Data sites: www.marad.dot.gov/library_landing_page/data_and_statistics/Data_and_Statistics.htm; www.bea.gov/international/index.htm;www.ndc.iwr.usace.army.mil/; www.dhs.gov/xlibrary/assets/statistics/yearbook/2008/ois_yb_2008.pdf

Figure 1 categorises the data according to its source. However, we could take an alternative approach which collects the data based on the type of decision that is to be made. We developed a systemic model which we call Port-Decision System Approach (PDSA) [3] which includes a number of subsystems to describe the seaport domain. Economic (ES) – shaded dark blue, Factors of productions and technology (FPT) – shaded brown, Global and environmental processes (GEP) – shaded green, Preference and experience (PE) – shaded skintone, Population and social structure (PSE) – shaded light blue and Political system institutions (PSI) – shaded purple. To make decisions concerning each of these subsystems it is necessary to extract the data from different sources and aggregate it by subsystem, shown for example in Figure 2.

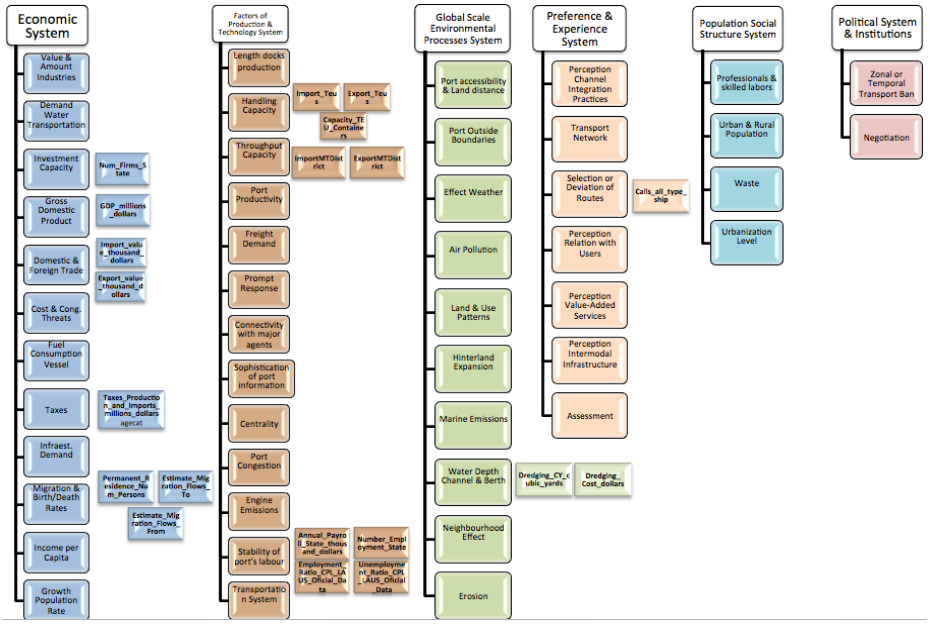

Fig. 2. Micro-level data repositories on the US Data websites

Looking closely at Figure 2 we can identify many low-level variables that have been compiled from multiple sources from the US data websites. Currently the decision maker is not necessarily aware that multiple hierarchies of data exist and would typically not have the skills or resources to combine the repositories to analyse the hierarchies. Our study involves exploration of these heterogeneous repositories in the quest for integrating data for analysis using data mining techniques so that evidence based guidance is provided for decision making.

Finding a way to connect macro and micro-level data will be important to aid strategic decision making. Strategic decisions, such as whether to expand the workforce, tend to concern macro level goals and data. However, data tends to be captured at the micro or operational level, such as number of employees and turnover rates. As a result there may be a mismatch between using micro level data for macro level decision making. On the one hand, it can be argued that the greater the level of abstraction of concepts represented in the model the more comprehensive the approach and

widely applicable the model will be to the phenomenon under study. However, a detailed representation of the model involving low level concepts (even instances) enhances its interpretability when implementing its outcomes in the real world. Table 1 shows how different level variables can map to subsystems and one another. In the next subsection we consider approaches in the literature and an approach using graph theory.

2.1 Data Aggregation Approaches in the Literature

There are techniques from the management field that consider how to handle the problem of data aggregation for decision making. Three of these techniques are: 1) multi-attribute value theory (MAVT), 2) aggregation of information based on indicators and 3) data level aggregation based on modelling abstraction. As described and used in [10], in this approach the attributes are associated to sub-attributes using expert weights ($W_{i,j}$) and an additive value function $Vc(a_j)$ that values an score between the preference of association and the given weight, illustrated in Figure 3.

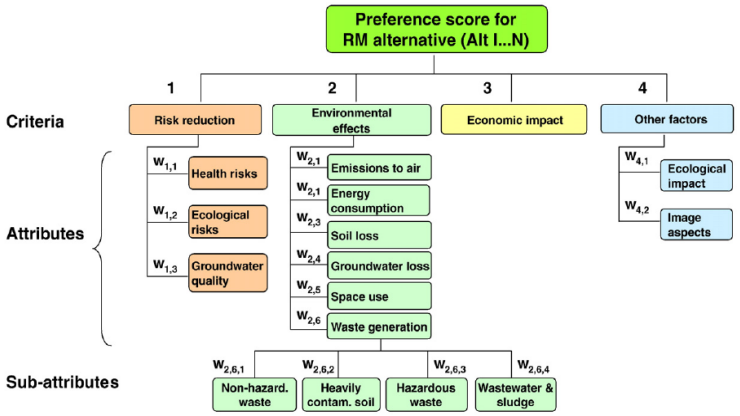

Fig. 3. Hierarchy of data aggregation based on MAVT [10]

A second approach on aggregation of information based on indicators has its roots in economic studies in which successive aggregation of scores are formed from different levels of indexes and sub-indexes. The 2011 World Economic Forum in their Global Competitiveness Report [12] uses this concept to report a structured computation of information. Formally, each sub-index represents a lower factor which can be measured from a data sample. The index is the weighted average of two or more sub-indexes. Finally, an indicator provides the higher factor which corresponds to an indication of the index worst and best possible outcomes. A third approach corresponds to typical data structures. Borshchev & Filippov [2] state that in general, aggregate values are used to model higher abstraction problems such as transportation networks. A decrease in the aggregation is performed when modelling problems use data to model exact sizes, distances, velocities and timings matter, as illustrated in Figure 4.

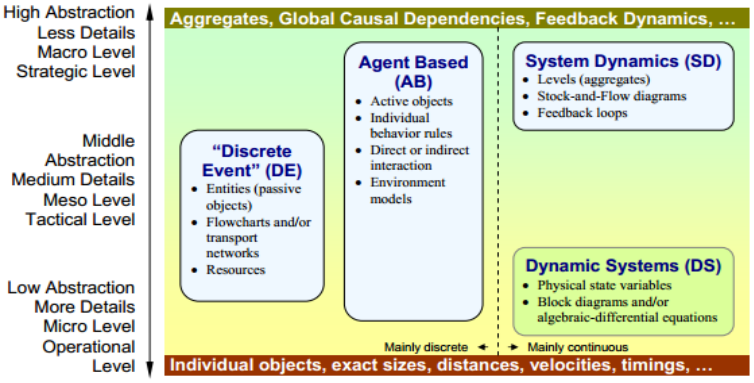

Fig. 4. Approaches in modelling according to the data level abstraction [2]

2.2 A Data Aggregation Approach Using Graph Notation

The previous approaches provide alternative solutions for data aggregation at different abstraction levels. Here to handle these different levels we suggest the use of graph theory to deal with hierarchical data structures. Graphs also provide visual benefits (Figure 5 shows a configuration based on the Table 1 formalisation) for understanding complex associations that otherwise need to be explained through complex analytical methods.

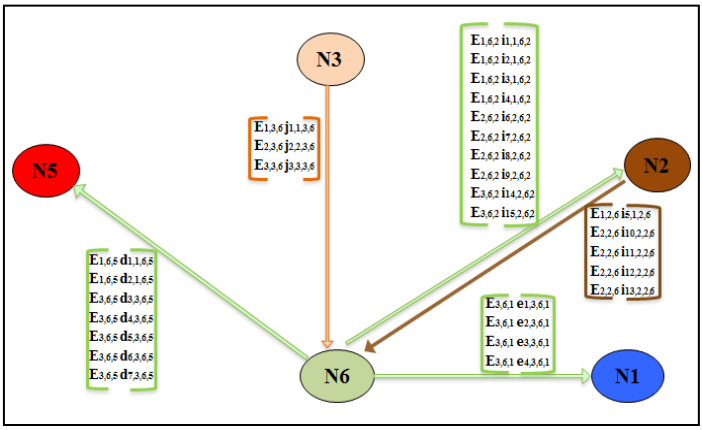

Fig. 5. Graph sample configuration

A mathematical definition of a graph G corresponds to a collection of vertices or nodes and edges that connect pairs of vertices. Suppose N denotes data at the macro level. This level aggregates concepts into categories that represent complex systems. $N_{i,j}$ is the pair of nodes denoting origin and destination of data in the macro level status, for example, $N_{6,2}$ traces a line from GEP (N_6) to FPT (N_2). E_k denotes the second-level data following the $N_{i,j}$ pathway, for example, $E_{1,6,2}$ denotes the concept for air pollution/emissions that relates with environment and production

systems. Finally, we can drill further down to find the micro-level data named here as edges: $a_{l,k,i,j}$ $b_{l,k,i,j}, c_{l,k,i,j}, d_{l,k,i,j}, e_{l,k,i,j}, f_{l,k,i,j}, g_{i,j}, h_{i,j}, i_{i,j}$, and $j_{i,j}$.

These edges display a cluster correlation of measurable variables which connect with the concepts described by the second-level data aggregation. We have been using data mining methods such as clustering and neural networks to identify relationships between variables and this work will be reported elsewhere.

Table 1. Formalisation - data hierarchies

Macro-level	N_i	Macro-level2	N_j	Second-level	$E\_(k,i,j)$	Micro-level	$a\_(l,k,i,j)$
GEP	N6	FPT	N2	Air pollution/emissions	E1,6,2	*CO2*	i1,1,6,2
						SO2	i2,1,6,2
						NOx	i3,1,6,2
						O3	i4,1,6,2
FPT	N2	GE	N6	Air pollution/emissions	E1,2,6	*facilities*	i5,1,2,6
PE	N3	GE	N6	Air pollution/emissions	E1,3,6	*Scientist*	j1,1,3,6
GEP	N6	PSI	N5	Air pollution/emissions	E1,6,5	*O3comply*	d1,1,6,5
						Inadequacies	d2,1,6,5
GEP	N6	FPT	N2	Water quality (Marine	E2,6,2	*oils*	i6,2,6,2
						chemicals	i7,2,6,2
						runoff	i8,2,6,2
						NMS	i9,2,6,2
FPT	N2	GEP	N6	Water quality (Marine	E2,2,6	*dredgeOcean*	i10,2,2,6
						needWtTreat	i11,2,2,6
						facilities	i12,2,2,6
						Inadequacies	i13,2,2,6
PE	N3	GE	N6	Water quality (Marine	E2,3,6	*Scientist*	j2,2,3,6
GEP	N6	ES	N1	Impacts of growth (land use patterns)	E3,6,1	*CRP*	e1,3,6,1
						MarketVal	e2,3,6,1
						LeaseNum	e3,3,6,1
						LeaseAcress	e4,3,6,1
GEP	N6	PSI	N5	Impacts of growth (land use patterns)	E3,6,5	*GAPStatus1*	d3,3,6,5
						GAPStatus2	d4,3,6,5
						GAPStatus3	d5,3,6,5
						GAPStatus4	d6,3,6,5
						CountyArea	d7,3,6,5
GEP	N6	FPT	N2	Impacts of growth (land use patterns)	E3,6,2	*LandFarms*	i14,3,6,2
						dredgeOcean	i15,3,6,2
PE	N3	GE	N6	Impacts of growth (land use patterns)	E3,3,6	*Scientist*	j3,3,3,6

3 Handling Missing Values

Most data integration systems focus on data aggregation. This issue is exacerbated by the fact, there are missing values affecting the different levels of aggregation. They are incorporated in any of the representations obtained and their analysis is useful to

facilitate knowledge discovery [7]. We discuss in this section our missing values approach after first describing the problem context of the example we provide.

3.1 Knowledge Discovery for the Environmental Dimension of Seaports

Many developments in methodology for incomplete data settings have predominately done in statistics. These methods need to be widely utilized in practice and thus we pose the question of how to arise new issues on missing values when conveying questions that PAs might want to answer in their deeds and duties. In previous work, we have identified data of a port with whom they should partner based on their compliance with environmental standards. Such a partnership can deliver competitive advantages and improved risk management performance. To identify who is compliant within the context of environmental management system standards (EMS) we need to identify what variables will be relevant. Key environmental issues are summarized in Table 2. The variables cover three main areas: ***Reducing Air Pollution/Emissions*** including particulate matter (PM), nitrogen oxides (*NOx*), sulfur oxides (*SOx*). carbon dioxides (CO_2), nitrogen oxides (NO_x,)sulphur dioxides (SO_2) and ozone expressed (*O3);* ***Improving Water Quality:*** Dredging activities (*dredgeOcean),* species habitat creation (national marine sanctuaries (*NMS*)); ***Minimizing Impacts of Growth:*** *CountyArea*. See Appendix in other paper by this author in this proceedings for descriptions of these variables.

Table 2. Observational dataset for missing value analysis

Selected Variable	Reducing Air Emissions	Improving Water Quality	Minimizing Impacts of Growth
needWtTreat		X	
facilities	X	X	
oils		X	
chemicals		X	
Inadequacies	X	X	
CO2	X		
O3comply	X		
O3	X		
O3cont	X		
SO2	X		
NOx	X		
CRP			X
LandFarms			X
Scientist	X	X	X
MarketVal			X
GAPStatus1			X
GAPStatus2			X
GAPStatus3			X
GAPStatus4			X
runoff		X	
CountyArea			X
LeaseNum			X
LeaseAcres			X
NMS		X	
dredgeOcean		X	X

Because these data are not always available, development of a missing value procedure is convenient for addressing several concerns caused by incomplete data. "Incomplete data may reduce the precision of calculated statistics because there is less information than originally planned. Another concern is that the assumption behind many statistical procedures is based on complete cases" [11. p.1]"

Table 3. Univariate Statistics for environment dataset

Univariate Statistics

	N	Mean	Std. Deviation	Missing		No. of Extremes[a]	
				Count	Percent	Low	High
facilities	40	36.25	39.135	4	9.1	0	3
oils	40	153188.627	502043.6682	4	9.1	0	6
chemicals	40	3385.675	11996.3488	4	9.1	0	5
CO2	41	95060248.87	86657623.70	3	6.8	0	0
O3cont	41	77.4073	17.03626	3	6.8	4	3
SO2	41	224628.944	234729.5499	3	6.8	0	0
NOx	40	79626.03	67356.842	4	9.1	0	0
CRP	44	50.5568	23.01462	0	.0	0	6
LandFarms	44	111196.70	174025.981	0	.0	0	7
GAPStatus1	43	8191340.26	25835111.17	1	2.3	0	9
GAPStatus2	44	5205314.80	9915808.516	0	.0	0	9
GAPStatus3	44	12868766.57	21033110.76	0	.0	0	10
GAPStatus4	44	53293715.89	53579654.96	0	.0	4	9
runoff	44	331.2382	307.09730	0	.0	0	1
CountyArea	44	1770.34984	5120.920416	0	.0	0	4
LeaseNum	43	940.23	1775.410	1	2.3	0	5
LeaseAcres	43	5039253.37	9388547.853	1	2.3	0	5
dredgeOcean	40	8255772.28	10254649.39	4	9.1	0	2
LeaseArea	43			1	2.3		
needWtTreat	44			0	.0		
Inadequacies	40			4	9.1		
O3comply	41			3	6.8		
O3	41			3	6.8		
Scientist	44			0	.0		
MarketVal	44			0	.0		
NMS	43			1	2.3		

a. Number of cases outside the range (Q1 - 1.5*IQR, Q3 + 1.5*IQR).

3.2 Missing Value Analysis

In this section we want to consider the impact of missing covariate data in the analysis of data aggregation at different abstraction levels. Horton and Switzer [4] report in a review of missing data methods from 26 original articles, how infrequent a missing covariate data analysis (i.e. multiple imputation) appears in observational studies. The impact of missing values is embedded in the data structure and therefore its analysis is critical. Typically, the methodology of missing covariate data answers the following questions:

1. Where are the missing values located?
2. How extensive are they?
3. Do pairs of variables tend to have values missing in multiple cases?

4. Are data values extreme?
5. Are values missing randomly?

Table 3 displays a summary of missing values for the sample of variables considered. We see that some values are not missing at all, while other variables, such as *facilities* and *oils*, are missing around 9% of the time. We conveniently assessed the most common methods (i.e. listwise, pairwise, regression estimation) with the assumption that the pattern of missing values does not depend on the data values, i.e. the data is missing completely at random (MCAR). However, running Little's [6] missing value test we conclude that significance value is less than 0.05 for our dataset. In this case data are not MCAR and then we need to use expectation-maximization (EM) estimation. EM depends on the assumption that the pattern of missing data is related to the observed data only (see Table 4). The overall summary of missing values is displayed in Figure 6 in three pie charts that show different aspects of missing values in the data. a) The variables chart shows that 14 of 24 variables have at least one missing value on a case. b) The cases chart shows that 11 of 44 cases have at least one missing value on a variable. c) The values chart shows that 40 of 1,056 values (cases x variables) are missing.

Table 4. Little's MCAR test, EM means : Little's MCAR test: Chi-Square=76.849, DF=56, Sig. = 0.34, The EM Algorithm failed to converge in 25 iterations

facilities	**Oils**	**chemicals**	**CO2**	**O3cont**	**SO2**
36.91	148940.68	3026.39	86651799.6	83.77	171032.42
NOx	**GAPStatus1**	**LeaseNum**	**LeasesAcres**	**DredgeOcean**	
63615.32	8192467.23	928.59	4976108.6	7062010.38	

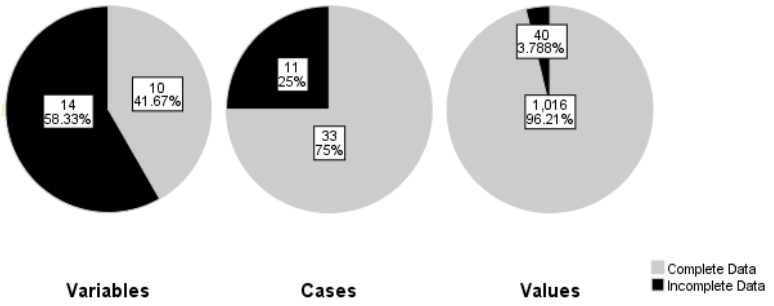

Fig. 6. Pie charts of summary of missing values

Table 5 indicates that three groups of variables record similar or related information: group 1 (*leaseNum*, *leaseAcres*, *NMS*), group 2 (*chemicals*, *oils*, *facilities*, *inadequacies*) and group 3 (*O3cont*, *SO2*, *CO2*, *Ocomply*, *NOx*). The table suggests that if we do not know the value of one variable within a group, probably we do not know the value for the other groups either.

The patterns chart in Figure 7 displays missing value patterns for the analysis variables. Each pattern corresponds to a group of cases with the same pattern of

Table 5. Patterns of missing data showing three groups. [a]Variables are sorted on missing patterns. [b]No of complete cases if variables missing in that pattern (marked with X) are not used.

	Missing Patterns[a]														
# Cases	GA[Status1	Lease-Num	LeaseA-cres	NMS	O3Cont	S02	C02	O3Comply	Nox	chem-icals	oil	facilities	Inadequ-acies	dred-geOcean	Comple-tIf
23															33
1									X						34
2														X	35
1	X													X	36
1										X	X	X	X	X	39
3										X	X	X	X		36
2					X	X	X	X	X						36
1		X	X	X	X	X	X	X	X						37

incomplete and complete data. For instance, pattern 4 represents cases that have missing values on group 3 (*O3cont*, *O3comply*, *O3*, *CO2*, *SO2*, *NOx*). The chart orders analysis and patterns to reveal where monotonicity exists. That is, there will be no "islands" of non-missing cells in the lower right portion of the chart and no "islands" of missing cells in the upper left portion of the chart. This dataset is nonmonotone and there are any values that would need to be imputed in order to achieve monotonicity.

The bar chart in Figure 8 shows that the majority of the cases in the dataset have pattern 1, i.e. the pattern for cases with no missing values. Patterns 2 and 4 represent missing values in around 5% of the cases. i.e., group 2 (chemicals, facilities, inadequacies, oils) and group 3 (O3cont, O3comply, O3, CO2, SO2, NOx) and pattern 6 that includes the variable dredgeOcean.

Estimated means are displayed in Table 5 for:

- The means from listwise deletion tend to be higher for group1 and group 2 whilst the means for chemicals, *CO2*, *CRP*, *GAPstatus1*, *GAPstatus3* and *LeaseNum* vary greatly. Because the data are not missing completely at random, estimates other than EM may be biased.
- The estimates for groups 2 and 3 with the greatest number of missing values include a large number of extreme values.

To observe if the distribution is more in line with the original data avoiding greater differences and random variations, it might be necessary to test the data to determine whether these values are not missing at random (MAR). Figure 9 displays multiple pairs of line charts, showing the mean and standard deviation of the imputed values of the variables chosen by the model as dependent at each iteration method for each of the 5 requested imputations. There should not be any patterns in the lines and look suitably random [10]. We see patterns that suggest the missing values are not random.

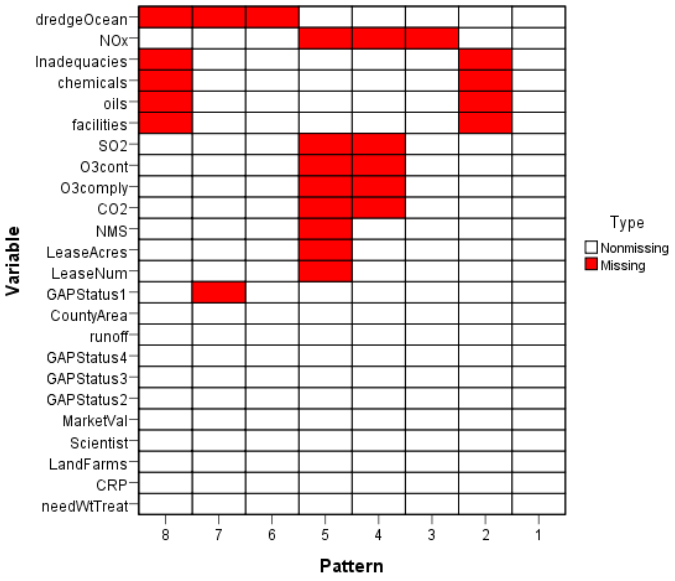

Fig. 7. Missing value patterns chart

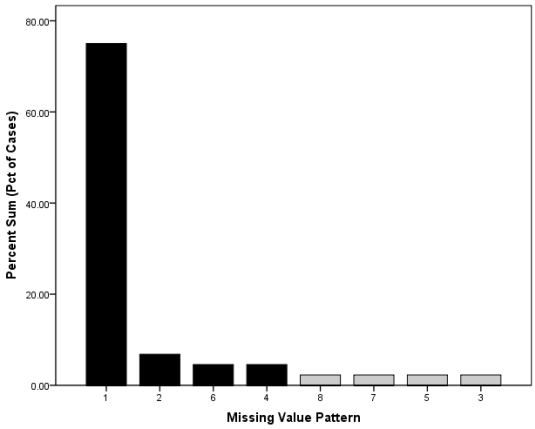

Fig. 8. Bar chart of missing value patterns

4 Conclusions and Future Work

Discovering knowledge from data for decision making is dependent on the existence of data relevant to the decision at hand. In the context of with whom PAs should partner based on their compliance with environmental management system standards (EMS), we have dealt with the maximum information from multiple levels and types of data, starting with macro-level data and ending with the micro-level data analysis.

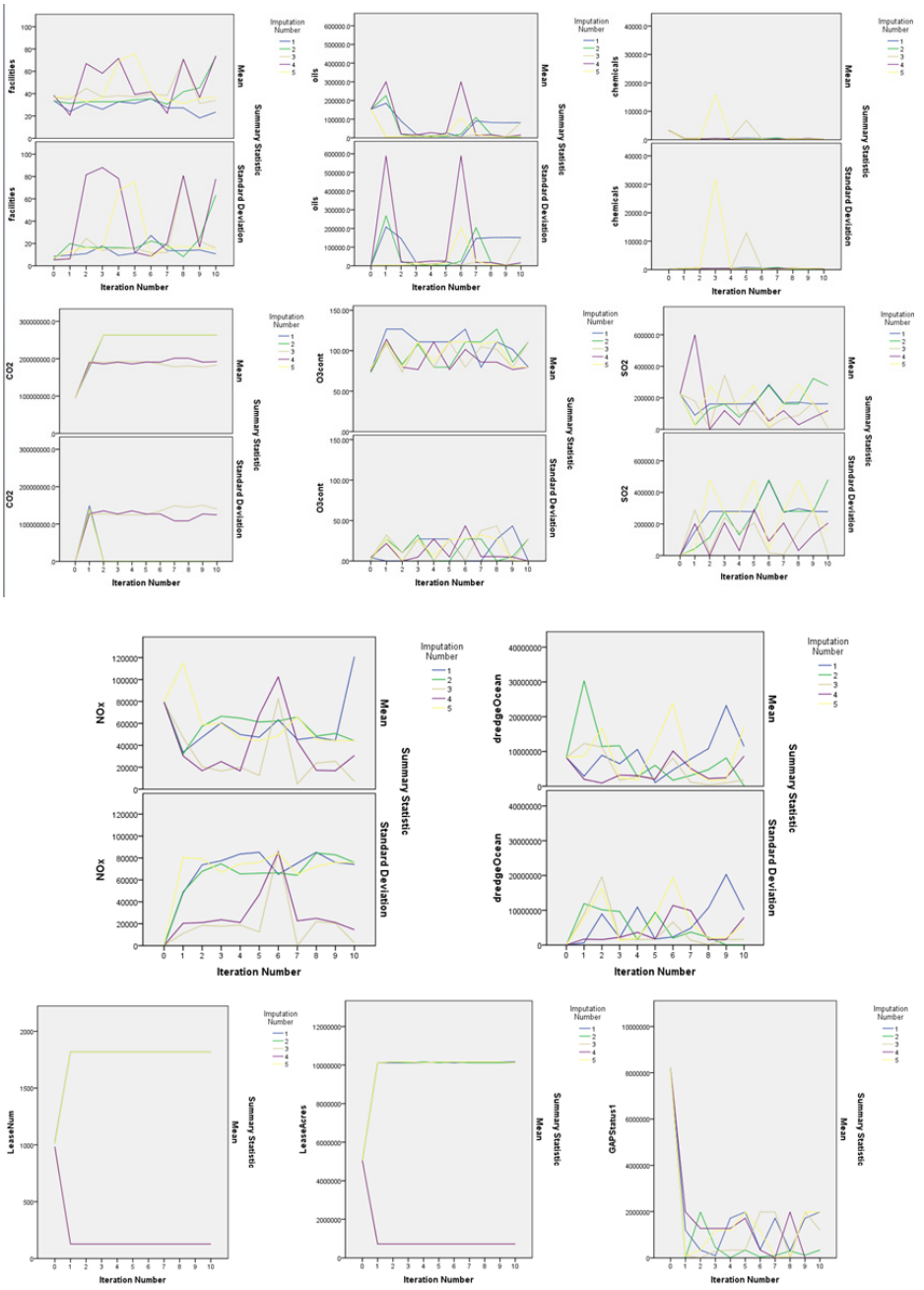

Fig. 9. Line charts to check if any patterns and that missing data are random

We are exploring the implementation of literature approaches on data aggregation such as [10], using graph notation to deal with hierarchical data structures and providing the visual benefits of graphs for understanding complex associations that otherwise need to be explained through complex analytical methods.

Missing value analysis suggests that if we do not know the value of one variable within a group, probably we do not know the value for the other groups either. The latter is corroborated our observations in that that dependency can be evident on variables pertaining to the same second-level of aggregation. That is, within the sample there is a correspondence of groupings displayed in the formalisation aggregation and the missing value pattern instances.

We will be conducting further analysis of the PDSA using time series data in a more comprehensive dataset for Latin American seaports in the quest to identify the legal, technical and political factors and associations that affect the decision making process of regional port authorities.

References

1. Bichou, K., Gray, R.: A critical review of conventional terminology for classifying seaports. Transportation Research Part A 39, 75–92 (2004)
2. Borshchev, A., Filippov, A.: From System Dynamics and Discrete Event to Practical Agent Based Modeling: Reasons, Techniques, Tools. In: The 22nd International Conference of the System Dynamics Society, Oxford, England (2004)
3. Halabi, A., Richards, D., Bilgin, A.: Proposing a port decision system approach for dynamic integration of South American sea ports. Paper Presented at the International Conference on Advances in ICT for Emerging Regions, ICTer (2011)
4. Horton, N.J., Switzer, S.S.: Statistical methods in the Jounal. New England Journal of Medicine 353(13), 1977–1979 (2005)
5. Kruse, C.J.: Environmental Management Systems at Ports - A new initiative. In: Proceedings of the 14th Biennial Coastal Zone Conference (2005)
6. Little, R.J.A.: A Test of Missing Completely at Random for Multivariate Data with Missing Values. Journal of the American Statistical Association 83, 1198–1202 (1988)
7. McClean, S., Scotney, B.: Using evidence theory for the integration of distributed databases. Int. J. Intell. Syst. 12, 763–776 (1997)
8. Notteboom, T.E.: Concentration and the formation of multi-port gateway regions in the European container port system: an update. Journal of Transport Geography 18(4), 567–583 (2010)
9. Notteboom, T.E., Rodrigue, J.-P.: Port regionalization towards a new phase in port development. Maritime Policy and Management 32(3), 297–313 (2005)
10. Sorvari, J., Seppälä, J.: A decision support tool to prioritize risk management options for contaminated sites. Science of the Total Environment 408, 1786–1799 (2010)
11. SPSS, Missing Value Analysis 16.0, Chicago, USA (2007)
12. World Economic Forum, The Global Competitiveness Report, Geneva, Switzerland (2011)