A System of Systems Approach to Analytics using Nolans Peer Relativity Profiling Transformations

Anthony G Nolan (G3N1U5), **Mark Johnson** (The Aerospace Corporation)**, Graham Williams** (Togaware)**, Emily Nolan** (G3N1U5), **Peter Phillips** (G3N1U5),

ABSTRACT

Nolan's peer relativity profiling transformations take each observation and assign it a value relative to all other members of the given sub-population. This allows for an observation on different measurement scales to be assigned a measurement on equivalent scales. Once combined with fuzzy logic, nth dimensional mathematics, and bounded within a framework it enables a system of systems analysis approach. In this paper, the methodology is applied to weather pattern analysis, which examines the total weather behaviour based on various systems like wind, sunshine, and air pressure to develop a set of rules for predicting rain, as well as other metrological events, which are associated with significant risk or economic loss. It is hoped that this research will lead towards a better understanding and development of warning systems for weather patterns, complex machine operations and other human safety and environment related system of systems.

Keywords: System of Systems, Nolan's transformations, pattern analysis, data mining, analytic(s), togaware.

1. INTRODUCTION

Attempting to understand large-scale, complex, systems of systems for predictive purposes is perhaps one of the most challenging endeavours undertaken by mankind. These endeavours encompass every arena of interest to humankind. As we learn more and more about our surroundings and our interactions with them, the more complete and intricate become our understandings of ourselves and our world and universe. In an effort to predict the influences and consequences of internal and external events, as well as, our interactions with our environments and world, we attempt to model the environments and actions versus reactions. The following paper and related work present one approach to modelling an environment through the applied use of nth dimensional mathematics, data mining, and soft computing techniques from a system of systems perspective.

This type of analytical process and modelling is not designed to achieve an outcome from a formula as in traditional data mining and analytical processes. The approach is more of a data pattern-matching process, which uses the pre-identification of a target group to be matched. This process requires using the entire population each time the model is run on a fresh dataset. Einstein's theory of relativity and Brenda Dervin's work on situationality, affords the perspective of taking an individual observation and through the combination of series of nth dimensional measurements, measuring the relativity between other observations which are defined by sub-population groupings [11]. Applying Einstein's theory of relativity to each observation, gives a unique opportunity to sequence an observation against all its peers in a number of relevant subpopulations or cohorts. By establishing its position in a number of cohorts which are then sequenced in alignment, this produces a DNA like chain of values, which are not affected by skewness nor appear to be affected by the curse of dimensionality. Because the process produces a new value for that observation in a standardised numerical scale, this allows for a between/within group analysis, which can be applied to a parent - child hierarchy; thus, achieving a system of systems approach.

Modelling large-scale, complex weather systems is used to present a system of systems approach to using Nolan's transformations. The process begins by assigning a cluster number to a complex pattern which is derived from a specific weather event. This is based on the cause and effect of the combinations of various weather factors, observations and models, which is then used to develop a recognisable sequential pattern

to profile the factors leading up to a weather event, and to establish a weather risk prediction sequence for a specific event. An example using a weather system of systems is preceded by implementation and background discussions.

2. BACKGROUND

Weather is key factor in our daily decision making that affects us in the areas of our leisure activities, dress and fashion coverings, transportation and communications; and impacts on shopping, working, health, and even survival. By definition the weather is the current and predicted near future state of the atmosphere for a particular location. In practicable terms, weather is the term given to describe a series of environmental systems characterized by variables and measurements, which can be categorized according to their impact on how we conduct our human activities. For instance; the level of temperature, the speed and direction of the wind, the amount of light, the chance of getting wet, and the amount of solar radiation we receive at any time; influence our behaviours, such as: how we dress, which mode of transportation to use, what time we undertake certain activities, and the risk to our completion of our activities in a safe and desired outcome.

2.1 Hotspots

A hotspot is a data driven approach to identifying a collection of observations that are of interest, including both those known to be of interest, and unknowns that have some "connection" to the known [1,2,3,4]. The hot spots idea is based on the work of Williams and Huang [4]. The idea is to cluster numeric data, and to then identify a cluster (or group of clusters) which, as a group, is interesting (e.g., the group has a high number of known risk observations, events, or large dollars at risk, etc.). We then build a decision tree, converted to rules, to describe the group (the "hot spot"). The group will generally also contain observations with an unknown status. The hypothesis is that these observations may be worthy of further investigation.

2.2 System of Systems

A system of systems in the context of the analytic process being pursued is a collection of systems that pool their resources and capabilities together to obtain a new, more complex, 'meta-system' which offers more functionality and performance than simply the sum of the constituent systems [9]. In analytics, a system of systems approach is really a system of model systems, where each model system is able to demonstrate or predict independently and interact with the other models comprising the analytic system of systems as needed. However, some product of that model system can contribute to other model systems at the same level or higher level models for the given analysis. Thus, each model is also part of a greater system model, for which the system model which is nearest to the principal question being asked can be called a Capstone Model (see Figure 1). A capstone model is dynamic, so a small change in a lower model can have a cascading effect to the parent models.

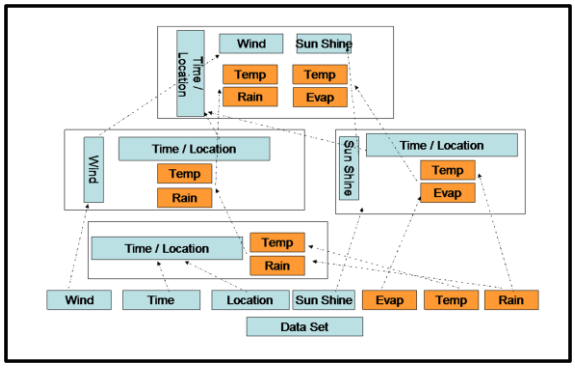

Figure 1. Capstone Model of an Analytic System of Systems

By using weather data as an example of a systems of systems approach, factors like wind, air pressure, sun shine hours, cloud cover, rain and evaporation, can be combined with other datasets from other observations like different locations, wave height, seasonal variations, sun solar effects and gravity. Hence, a very complex model can be developed to analyze trends in weather events, which is a capstone approach to the contributing systems.

2.3 Nolan's Peer Relativity Profiling Transformations

Nolan's Peer relativity profiling transformations methodology is about taking an observation, which is defined by a series of characteristics which defines a subpopulation. Each subpopulation is extracted from the entire population, and the observation is then assigned a value which is relative to all the other members of that sub-population. For example, if an observation has five different characteristics to describe it, then that observation exists in five different subpopulations, and hence that value will be represented five different times, each representation ranging with values between 0 and 99. More detailed information concerning these processes can be found in the following references [5,6].

3. IMPLEMENTATION

3.1 The Process

The flow chart, shown in Figure 2, is the basic process of the model. Once the software application is started, the data is loaded, transformed, and then the k-means and clustering processes are repeated until the cycle is completed. The software then locates the hotspot cluster, calculated by a series of performance indicators. These are derived by the size of the cluster, the number of target observations and the strike rate. A list is generated (see Figure 3) which lists the different key performance indicators. The software application allows for the routine to be rerun, with the option to choose clusters rather than the software application.

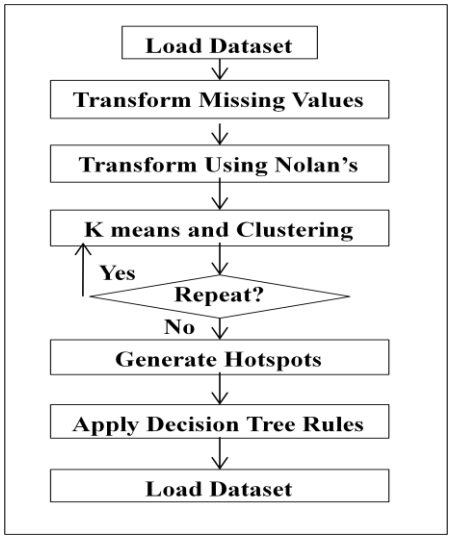

Figure 3. Process Flow Chart

At the end of the process the software generates a data project file which can be loaded into another data mining package called rattle. Rattle (the R Analytical Tool To Learn Easily) is a data mining toolkit used to analyze very large collections of data. Rattle presents statistical and visual summaries of data, transforms data into forms that can be readily modeled, builds both unsupervised and supervised models from the data, presents the performance of models graphically, and scores new datasets. Through a simple and logical graphical user interface based on Gnome [7], Rattle can be used by itself to deliver data mining projects. Rattle also provides an entry into sophisticated data mining using the open source and free statistical language R [8]. Rattle runs under GNU/Linux, Macintosh OS/X, and MS/Windows. The aim is to provide an intuitive interface that takes you through the basic steps of data mining, as well as illustrating the R code that is used to achieve this. Whilst the tool itself may be sufficient for all of a user's needs, it also provides a stepping stone to more sophisticated processing and modeling in R itself, for sophisticated and unconstrained data mining [10]. An example output is shown in Figure 4.

Clusters:

	group	size	sum	nonc	comp	unkn	strike	risk	absrisk	averisk
1	1	23	8	9	1	13	39.13	0k	0k	0k
2	2	8	8	8	0	0	100	0k	0k	0k
3	3	14	3	3	0	11	21.43	0k	0k	0k
4	4	6	1	2	1	3	33.33	0k	0k	0k
5	5	7	1	1	0	6	14.29	0k	0k	0k
6	6	9	0	1	1	7	11.11	0k	0k	0k
7	7	14	9	10	1	3	71.43	0k	0k	0k
8	8	5	-2	0	2	3	0	0k	0k	NaNk
9	9	8	-1	0	1	7	0	0k	0k	NaNk
10	10	11	2	2	0	9	18.18	0k	0k	0k
11	11	7	3	3	0	4	42.86	0k	0k	0k
12	12	11	6	7	1	3	63.64	0k	0k	0k
13	13	4	3	3	0	1	75	0k	0k	0k
14	14	12	9	9	0	3	75	0k	0k	0k
15	15	5	5	5	0	0	100	0k	0k	0k
16	16	10	-2	0	2	8	0	0k	0k	NaNk

Figure 2. Example Hotspots List

4. EXAMPLE APPLICATION

4.1 The Data

The weather dataset comes from data provided from the Australian Bureau of Metrology, and are the observations of Canberra. Canberra is the capital of Australia, and is located at Latitude: 35.30 °S

Longitude: 149.20 °E and is around 578 meters above sea level. Canberra is on the east coast of Australia approximately 150 kilometers inland from the coast. It is considered to have a fairly dry Continental Climate. Canberra has a mountain range called the Brindabellas, which has a mild protective effect and creates a rain shadow in some areas of Canberra. Also, Canberra has a high percentage of parks and green zones, intermixed with urban development, which produces pockets of different weather behavior within small distances of each other. The temperature ranges between -5 and 21 degrees Celsius. Rainfall ranges from zero to 40 mms of rain, with a total of 522 mms of rain for the year for this dataset, and rain occurred on 103 days.

In Figure 4, one can see a decision tree model and a subset of the rules which were used to derive the decision tree. These rules are derived from the analysis of the kmeans derived hotspots using Nolan's Transforms (with clustering), as the new target group; however, the rule generation is based on the original data observations. Table 1 presents the weather dataset variables.

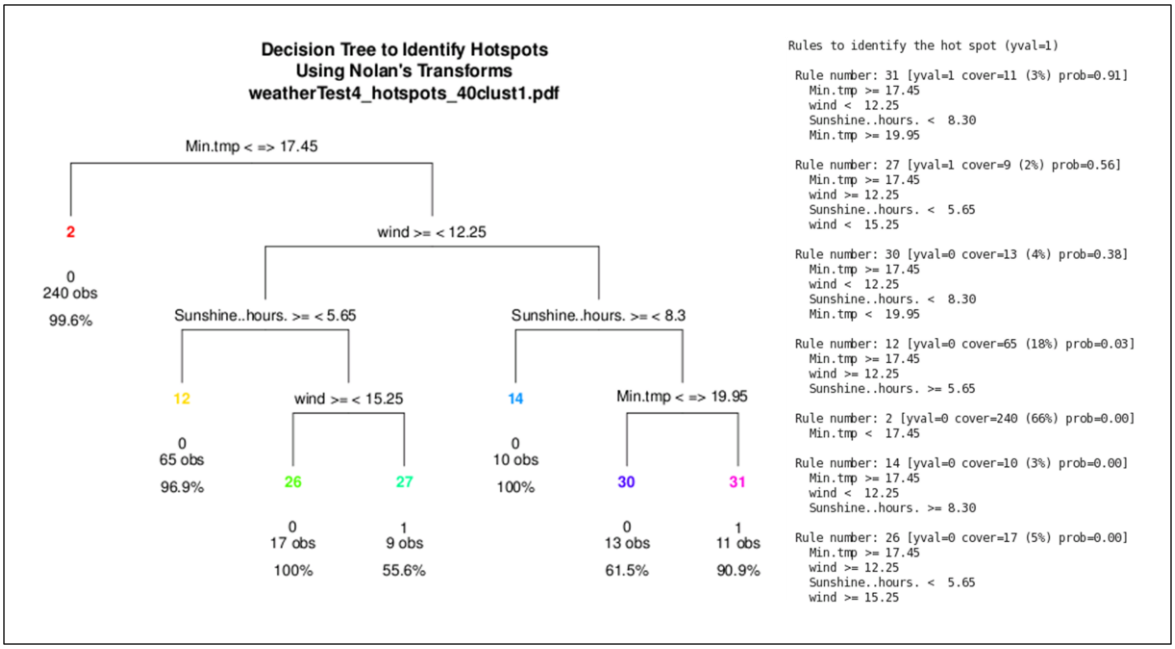

Figure 4. Example Decision Tree and Rules Subset Derived Using Nolan's Transforms

Heading	Meaning	Units
Date	Day of the month	Day
Location	Location of observations	Name
Min Temps	Minimum temperature in the 24 hours to 9am.	degrees Celsius
Max Temp	Maximum temperaure in the 24 hours from 9am.	degrees Celsius
Rainfall	Precipitation (rainfall) in the 24 hours to 9am.	millimeters
Evaporation	Class A pan evaporation in the 24 hours to 9am	millimeters
Sunshine	Bright sunshine in the 24 hours to midnight	hours
WindGustDir	Direction of strongest gust in the 24 hours to midnight	16 compass points
WindGustSpeed	Speed of strongest wind gust in the 24 hours to midnight	kilometers per hour
WindSpeed	Wind speed averaged over 10 minutes intervals	kilometers per hour
Humidity	Relative humidity	percent
Pressure	Atmospheric pressure reduced to mean sea level (MSL)	hectopascals
Cloud	Fraction of sky obscured by cloud	eighths
RainToday	Did it rain the day of the observation	Yes/No

Table 1. Variables

A number of extra variables were also derived as change variables, which were calculated from data observed at 9.00 am and 3.00 pm to establish the direction and degree of movement of the specific variables, to establish a pattern of movement.

Table 2 is an example of the results of using Nolan's Peer Relativity Profiling Transformations, which shows the target cluster, and a sample segment of the data matrix of minimum temperature and its relative position in each sub cohort. For example, looking at day 21, one can see that the real minimum temperature was 17.7, which was a 2.2 degrees C increase from the previous observation (in sequence). Within the cohorts, its score was in the 82nd percentile for overall minimum temperature, 73rd percentile for humidity, 70th for wind direction, 76th for wind speed, 79th of air pressure and 71st for cloud cover. Thus, one can see that relative to all other observations in the entire dataset, there is a difference between that observations position. Also, when looking at that days observation against all the others in the hotspots cluster, it is different to the others, but not significantly, where as days 230 and 319 are a lot more alike.

The target group is defined by the amount of rain, and the amount of evaporation over 1 years of observations.

					Min.tmp					
Day	**Tgt**	**Risk**	**Temp**	**Min.tmp**	**Temp_**	**Rel.Hum**	**Wind.Dir**	**Speed**	**MSL**	**Oct**
19	1	0.2	2.7	19.2	76	73	98	87	99	82
21	1	22.4	2.2	17.7	82	73	70	76	79	71
32	0	0	0.9	21	72	84	89	99	81	99
51	0	0	-0.7	22.9	99	99	98	99	99	99
54	1	4.4	2.9	20.5	89	80	85	85	99	82
55	0	0	3.4	22	99	99	94	87	99	99
58	0	0	6.1	21.7	99	99	92	99	99	99
67	0	0	7	23	99	99	99	99	99	99
79	1	0.6	1.6	17.5	59	0	82	29	80	68
97	1	12	0.5	21.7	99	99	98	99	99	93
135	0	0	0.9	22.2	90	93	99	99	94	99
187	0	0	8.4	10.6	99	28	31	22	99	30
193	1	1	5.8	20	99	82	82	84	79	86
222	1	5.2	2.5	19.6	99	85	82	80	99	84
230	1	0.2	3.4	20.2	78	94	84	73	89	88
301	0	0	2.2	19.4	99	78	85	78	79	81
302	1	62.6	4.9	17.9	99	99	77	64	86	70
305	0	0	0.8	19.7	93	75	83	84	81	78
319	0	0	0.5	20.5	74	78	98	86	98	87
330	0	0	4	18.1	72	82	72	65	77	72
348	0	0	2.2	18.9	94	71	85	58	72	79
354	0	0	0.9	22.8	99	99	99	99	99	99
361	0	0	1.6	21.4	90	86	99	99	96	92

Table 2. Example Output from the System of Systems Analytic Process Using Nolan's Peer Relativity Profiling Transformations

5. SUMMARY

While using this technique on weather data is still in its early days, this technique has been successfully applied to financial and environmental profiling. Early runs have also shown a higher than random results in detecting patterns that lead up to weather events like cold snaps and rain. With the cost in economic

terms of Acts of God, Natural Disasters cost billions of dollars each year. Even milder weather changes cost agricultural producers in the USA over $1 billion US per year. Early detection of adverse weather events, can save money and prevent human deaths. The next stage being undertaken is to combine other datasets from more locations, as well as other environment factors to add to the modeling process. We envisage that once a capstone model approach is applied to multiple locations, then pre-warning hotspots identification which creates a rule base warning system, will contribute to the risk reduction of weather damage.

If you would like to know more, you can contact Tony Nolan, tony@g3n1u5.com; Mark Johnson, mark.a.johnson@aero.org; Graham Williams, Graham.Williams@togaware.com; Emily Nolan, Emily@g3n1u5.com; or Peter Phillips, pxphill@gmail.com.

REFERENCES

[1] Denny, Graham Williams and Peter Christen, Exploratory Hot Spot Profile Analysis using Interactive Visual Drill-Down Self-Organizing Map, In Proceedings of the 12th Pacific Asia Conference on Knowledge Discovery and Data Mining (PAKDD08), Osaka, Japan, May 2008. Lecture Notes in Computer Science.

[2] Denny, Graham Williams and Peter Christen, Exploratory multilevel hot spot analysis: Australian Taxation Office case study. In Proceedings of the Australasian Data Mining Conference, CRPIT Volume 70 (AusDM07), Gold Coast, Australia, December 2007.

[3] Graham Williams, Evolutionary Hot Spots Data Mining, In Proceedings of the 3rd Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD99), Beijing, China, April 1999.

[4] Graham Williams and Zhexue Huang, Mining the Knowledge Mine: The Hot Spots Methodology for Mining Large Real World Databases; Advanced Topics in Artificial Intelligence Lecture Notes in Artificial Intelligence, Volume 1342, Pages 340—348, Springer, 1997.

[5] A. Nolan, Getting underneath the decision making process, In Modsim 95 proceedings, Modelling and Simulation Association, Newcastle, 1995.

[6] A. Nolan, Modelling and Applications of Multi-Dimensional Interval Data to Artificial Life, Control Systems, Decision Making, In Proc. Fourth World Automation Congress, pp. 345-360, Hawaii, USA 2000.

[7] http://en.wikipedia.org/wiki/GNOME

[8] http://en.wikipedia.org/wiki/R_programming_language

[9] C. Keating, Critical Challenges in the Maturation of the System of Systems Engineering Paradigm, In Proc First IEEE Conference on System of Systems, Los Angeles, CA, 2006.

[10] http://www.togaware.com

[11] C. D. Reinhard and B Dervin, The Application of Dervin's Sense-Making Methodology to Media Reception Studies: Interpretivism, Situationality, and the Empowerment of Media Users, ECREA Subdivision Conference, Transforming Audiences 2.0, London, Sept 2009.