DIAGNOSTIC ANALYTICS CASE STUDIES ANALYSES

This document consists of two sections that contain the following information:

1. Analyses of diagnostic analytics case studies.
2. Analysis of the results produced by the case study analysis.

1. Case Study Analyses

Case Study 1: Diagnostic Analysis for outlier detection in big data analysis [1]

Initiation:

Pro requirement 1.1: This project seeks to "addressed the concept of data quality diagnosis to identify the outlier presented in the dataset"

Gen requirement 1.1: The analytics project must have a clearly defined goal.

Pro requirement 1.2: "big data", "Data quality", and "Outlier" were defined within the context of the data analytics project.

Gen requirement 1.2: The analyst be aware of the level of expertise of the client and define the key terminology within the context of the data analytics project accordingly.

Pro requirement 1.3: This analytics project must undertake "Data quality diagnosis was run on the dataset to understand the data and identify errors that appeared in the dataset".

Gen requirement 1.3: The analytics project must have a clearly defined system/object on which the diagnostic analysis is carried out.

Pro requirement 1.4: This project will use "outlier" which is "evaluated by comparing them with the general distribution of the values inside the column" in order to evaluate the dataset..

Gen requirement 1.4: The analytics project must have a quantitative metric(s) that is used to evaluate the system.

Acquisition:

Pro requirement 1.5: The "Global Food Prices Dataset" will be obtained from "Humanitarian Data Exchange".

Gen requirement 1.5: The analytics project must have defined which data sets will be used and where these data sets will be acquired.

Pro requirement 1.6: The dataset “contains 1048576 records and 17 column listings which consist of the following attributes; Country Id, Country Name, State Id, State Name, Market Id, Market Name, Food Id, Food Name, Currency Id, Currency Name, Type Id, Type Name, UnitMetric Id, UnitMetric Name, Month, Year, price and Commodity Source”.

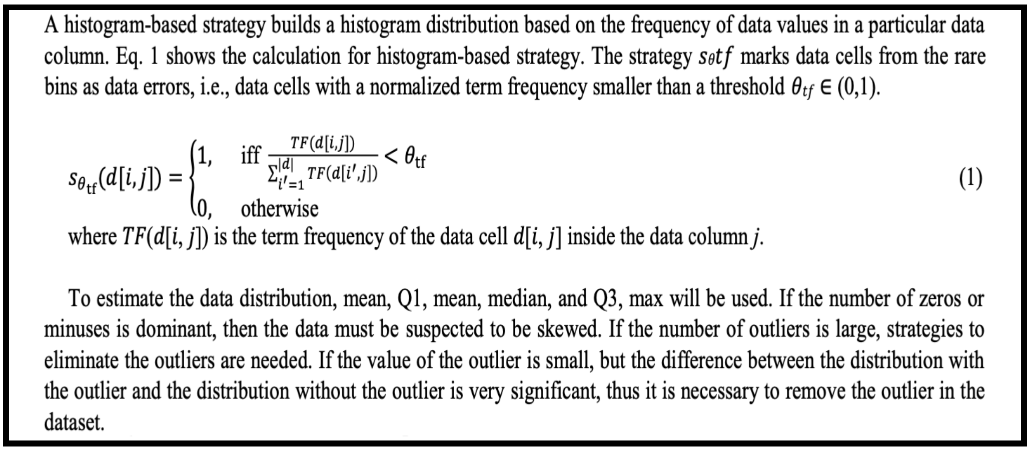
Gen requirement 1.6:The properties to the dataset that is used within the analytics project must be defined.

Analysis:

Pro requirement 1.7: In order to identify the “the outlier” “a histogram-based strategy is chosen”.

Gen requirement 1.7: The ‘strategy’ that will be employed to carry out the diagnostic analysis must be defined.

Pro requirement 1.8: The specifics regarding the histogram based strategy for this project is defined as:



A histogram-based strategy builds a histogram distribution based on the frequency of data values in a particular data column. Eq. 1 shows the calculation for histogram-based strategy. The strategy $s_{\theta}tf$ marks data cells from the rare bins as data errors, i.e., data cells with a normalized term frequency smaller than a threshold $\theta_{tf} \in (0,1)$.

$$s_{\theta_{\text{tf}}}(d[i,j]) = \begin{cases} 1, & \text{iff } \frac{TF(d[i,j])}{\sum_{i'=1}^{|d|} TF(d[i',j])} < \theta_{\text{tf}} \\ 0, & \text{otherwise} \end{cases} \qquad (1)$$

where $TF(d[i,j])$ is the term frequency of the data cell $d[i,j]$ inside the data column j.

To estimate the data distribution, mean, Q1, mean, median, and Q3, max will be used. If the number of zeros or minuses is dominant, then the data must be suspected to be skewed. If the number of outliers is large, strategies to eliminate the outliers are needed. If the value of the outlier is small, but the difference between the distribution with the outlier and the distribution without the outlier is very significant, thus it is necessary to remove the outlier in the dataset.

Fig. 1.1 An extract taken from the research article describing the ‘strategy’[1]

Gen requirement 1.8: An analytics project must have an in depth definition of the strategy that will be used to carry out the analytics. Which includes specific equations that will be used and what the variables within said equation are.

Presentation:

Pro requirement 1.9:This analytics project must result in a “visualization of the price in the dataset with the outlier and without the outlier” showing that the “presence of the outlier provides a significant difference in the graph”.

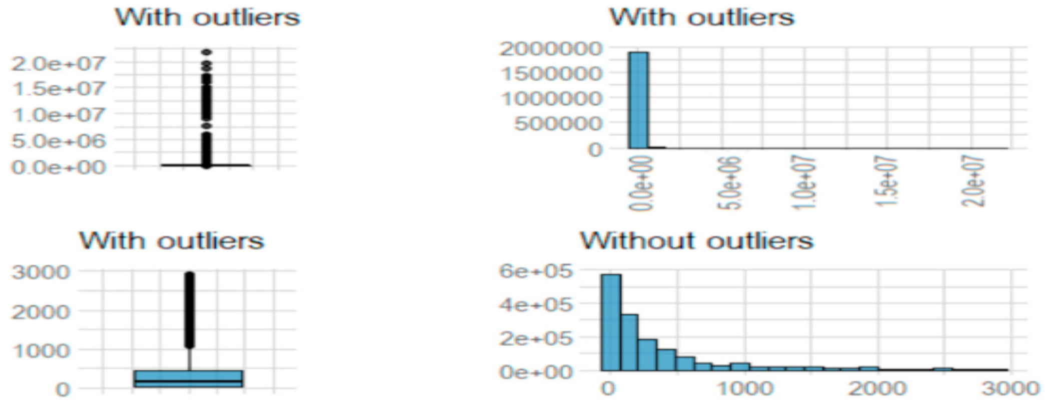


Fig. 2. Outliers in the Global Food Price dataset.

Fig. 1.2 An extract taken from the research article describing the 'strategy'[1]

Gen requirements 1.9: The graphical representations that are required when presenting the diagnostic results must be defined.

*Although this data analytics project is defined as a diagnostics project there are also aspects of prescriptive analytics given that the authors provide suggestions as to what should be done when an "outlier" is found within the data set.

Pro requirement 1.10: The results of the analytics projects must define the factors that are responsible for the "outlier" explicitly as follows "currency, year, prices, location and type of food".

Gen requirements 1.10: The format in which the results of a diagnostic project are textually presented must be defined.

Case study 2: Diagnostic analysis of regional ozone pollution in Yangtze River Delta, China: A case study in summer 2020 [2]

Initiation:

Pro requirement 1.1: The goal of this analytics project is to conduct "a comprehensive diagnostic analysis of O3 formation during a 1-week regional O3 pollution event in August 2020 in the YRD region".

Pro requirement 1.2: "Emission based model (WRF-CMAQ)", and "OBM" were defined within the context of the data analytics project.

Pro requirement 1.3: This analytics project "aims to understand the causes of O3 pollution during " "A regional ozone (O3) pollution event occurred in the Yangtze River Delta region during August 17–23, 2020 (except on August 21)".

Pro requirement 1.4: This project will use "O3 pollution" as well as "O3 sensitivity to its precursors during the O3 pollution" both of which will be measured using "O3 concentrations".

* This information is not clearly defined and within the introduction of the paper and this results in a level of ambiguity when defining the requirement.

Acquisition:

*This research paper is very unique because the authors predominantly rely on models as opposed to a dataset inorder to carry out the diagnostic analysis, although inputs required for one of the models was defined using a dataset. Therefore it is advisable to define both the datasets and models used within a diagnostic analytics project given that models and datasets are not interchangeable.

Pro requirement 2.1: The models used within this project are "Weather Research and Forecasting (WRF) model version 4.2.1 was used to provide the meteorological fields for the chemical transport model",

"emission-based model (EBM) (i.e., a 3-D chemical transport model) was used to simulate the air quality during the episode",

"The Community Multiscale Air Quality version 5.2 (CMAQv5.2), developed by the United States Environmental Protection Agency (US EPA), was employed in this study to simulate the air quality and explore the causes of O3 pollution during the summer 2020 in the YRD region" ,

"a source-oriented CMAQ model was utilized in this study to assess the contributions of different emissions sources and emitting regions to O3, which was based on an improved sensitivity regime classification (i.e., VOC-limited, NOx-limited, and transition regimes) approach for O3 formation",

"OBM developed by Cardelino and Chameides (1995), incorporating the Master Chemical Mechanism version 3.3.1 (MCMv3.3.1, available at http://mcm.leeds.ac.uk/MCM/) in this study, was used to simulate the O3 photochemistry and further identify the sensitivity of O3 formation to precursor concentrations at a certain monitoring site".

Gen requirement 2.1: The model(s) used within the analytics project along with what said models are used for must be defined for a diagnostic analytics project.

Pro requirement 1.5: The datasets used within this data analytics project are "1° × 1° FNL reanalysis dataset with a temporal resolution of 6 h from the National Centers for Environmental Prediction" , the "hourly observation data of trace gases (e.g., O3, NO2, and CO) for major cities in the YRD region were obtained from the China National Environmental

Monitoring Center (CNEMC, http://106.37.208.233: 20035/) from August 17 to 23, 2020.", "Continuous field measurements of VOCs were also carried out at a typical urban monitoring site (32.057°N, 118.749°E, Fig. 1(b)) in Nanjing that was surrounded by commercial and residential districts. Hourly data of 57 VOC species, consisting of 29 alkanes, 16 aromatics, 11 alkenes, and acetylene, were collected. The observed data of NO2, CO and VOCs were as input in OBM. The real-time hourly data of the meteorological parameters (i.e., temperature, wind speed, wind direction, relative humidity, and precipitation) in Nanjing were obtained from the weather website (http://q-weather.info/weather/)."

Pro requirement 1.6

Analysis:

Pro requirement 1.7: The analytics will be carried out using "an emission- based model" and "an emission- based model".

Pro requirement 1.8: Although the strategy used within this study was defined in an in-depth manner it is very extensive and hard to grasp, owing to the fact that the authors had a reader that is more well versed in the subject matter in mind when developing the research article. Although it can be said that the methodology could have been more concisely defined.

Pro requirement 2.2: Within this data analytics project "integrated process rate (IPR) module, a process analysis tool based on solving the mass continuity equation, was available in CMAQv5.2 and applied in this study".

Gen requirements 2.2: The analytics project must have clearly defined the tools that are going to be used and as well aa what those tools will be used to do.

Presentation:

Gen requirement 1.9: The results of this analytics project must be formatted such that "~~The OBM analysis determined that~~ the O3 formation was in the ~~VOC-limited regime on August 19,~~ and in the transition regime on all the other polluted days. ~~Although neither aromatics nor alkenes were~~ the most abundant groups, ~~they were~~ the top two contributors to O3 formation in terms of the shares in OFP among all the VOCs measured", and "The process analysis indicated that ~~the photochemical process~~ was the predominant factor in the formation and accumulation of O3 during the daytime."

Pro requirement 1.10: This data analytics project must visually represent:

“Modeling domain. (a) The three nested domains with different horizontal resolutions (d01: 36 km, d02: 12 km, and d03: 4 km) for WRF simulation. The blue rectangle indicates the CMAQ simulated domain. (b) The various colored and patterned areas represent 15 cities tagged in the YRD region. The black dot identifies the location of the VOCs field measurement.”

“Time series of observed concentrations (black dotted line) of (a) NO2 and (b) total VOCs; (c) comparison between simulated (red dotted line) and observed O3 concentrations (black dotted line); the daily average concentrations of (d) NO2, (e) total VOCs, and (f) MDA8 O3 (the red dash line for the limit exceeding 160 μg/m3) in Nanjing during study period.”

“Time series of O3 change rate caused by individual atmospheric processes in Nanjing in the PBL. DEPO, HTRA, VTRA, and CHEM mean deposition (dry deposition and cloud process), horizontal transport, vertical transport, and chemical process, respectively. Total O3 variation is the sum of these processes.”

“The RIR values for O3 precursors (i.e., AVOCs, NOx, BVOCs (isoprene), and CO) at Nanjing urban site during O3 pollution episode.”

“(a) RIR of top 10 AVOCs for O3 formation at Nanjing urban site, and (b) concentrations and (c) OFP proportions of different VOCs groups (ALKA: alkanes; ALKE: alkenes; AROM: aromatics; and ACET: acetylene) to total observed VOCs during O3 pollution episode.”

“The percentage of source contributions to average MDA8 O3 attributed to (a) power, (b) industry, (c) residential, (d) transportation, (e) biogenic source, (f) IC/BC, and (g) background during O3 pollution episode in the YRD region. Contributions of IC/BC to O3 are attributed to NOx and VOCs entering the domain through the initial and boundary conditions. Background O3 is regarded as that directly entering the domain through the initial and boundary conditions. The area scope of Nanjing city is marked in bold on the map.”

“The percentage contributions of different sources to hourly O3 in Nanjing from August 17–23 (excluding August 21), 2020. Predicted O3 concentrations from different sources are represented by the corresponding colored areas.”

“Source contributions of transport from individual cities to hourly O3 in Nanjing from August 17 to 23 (excluding August 21), 2020. “BG” means background. The percent in the pie chart is the average MDA8 O3 during O3 pollution episode. “Local” refers to the contribution of Nanjing city itself. “Non-Local” refers to the contribution from cities tagged other than Nanjing. “Other” refers to contributions from those cities not tagged in the target area.”

Note: This case study provided much insight into the complexity of diagnostic analytics when compared to descriptive analytics where although one cannot elicit requirements that capture the complexity of the methodology adequately, having the foundational requirements set as to what methods will be used will provide the data analysts a much-needed foundation to build upon doing his or her projects.

Case study 3: Mixed logit model based diagnostic analysis of bicycle-vehicle crashed at daytime and nighttime [3]

Initiation:

Pro requirement 1.1: The goal of this analytics "is to explore the underlying factors to injury severity in crashes involving cyclists in the daytime and nighttime separately".

Pro requirement 1.2: "Mixed logit (ML) model", and "Marginal effect analysis" must be explained within the context of the data analytics project.

Pro requirement 1.3: This data analytics project will look at "crashes involving cyclists in the daytime and nighttime separately"

Pro requirement 1.4: This project will use "Five injury severity levels are identified, which are no injury (NI), possible injury (PI), suspected minor injury (SMI), suspected severe injury (SSI), and fatal injury (FI)" in order to evaluate the severity of "crashes involving cyclists".

Acquisition:

Pro requirement 2.1: This analytics project will utilize "mixed logit model to analyze the underlying factors towards injury severities in crashes involving cyclists"

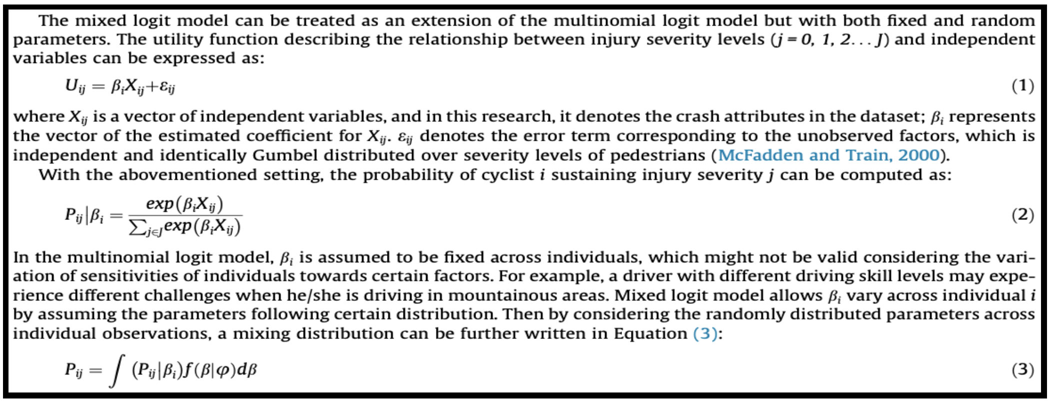
Pro requirement 1.5: Inorder to carry out the analysis for this project "data used to estimate mixed logit models are retrieved from the police report data of North Carolina Department of Transportation (NCDOT) between 2007 and 2018".

Pro requirement 1.6: The properties of the data "include many categorical explanatory variables, which are cyclist, driver, vehicle, road, environment, and crash characteristics. 8049 out of 11,196 are filtered for model estimation via the data cleaning process. Essentially, the data without the necessary information were filtered out in the cleaning process. The cyclist's characteristics contain the gender and age of the cyclists, as well as alcohol usage. The characteristics of drivers include the same variables as those of cyclists. Vehicle characteristics mainly refer to vehicle type. Variables in road characteristics are traffic control, speed limits, road configuration and road condition, rural and urban. Environmental char- acteristics include weather, light condition, region, and development type. Crash characteristics contain variables of crash types, crash time, and crash location. Five injury severity levels are identified, which are no injury (NI), possible injury (PI), suspected minor injury (SMI), suspected severe injury (SSI), and fatal injury (FI). In this study, no injury is selected as the base injury severity level in the mixed logit model. Details of the data utilized in this study are summarized in Table 2 by category and injury severity level."

Analysis:

Pro requirement 1.7: The analytics will be carried out using a "mixed logit model" as well as "Marginal effect analysis".

Pro requirement 1.8: The in-depth description for the methods used within this data analytics project for the mixed logit model and the marginal effect analysis are shown in **Figure 1.3** and **Figure 1.4** respectively.



The mixed logit model can be treated as an extension of the multinomial logit model but with both fixed and random parameters. The utility function describing the relationship between injury severity levels (*j* = *0, 1, 2*. . . *J*) and independent variables can be expressed as:

$$U_{ij} = \beta_i X_{ij} + \varepsilon_{ij} \quad (1)$$

where X_{ij} is a vector of independent variables, and in this research, it denotes the crash attributes in the dataset; β_i represents the vector of the estimated coefficient for X_{ij}. ε_{ij} denotes the error term corresponding to the unobserved factors, which is independent and identically Gumbel distributed over severity levels of pedestrians (McFadden and Train, 2000).

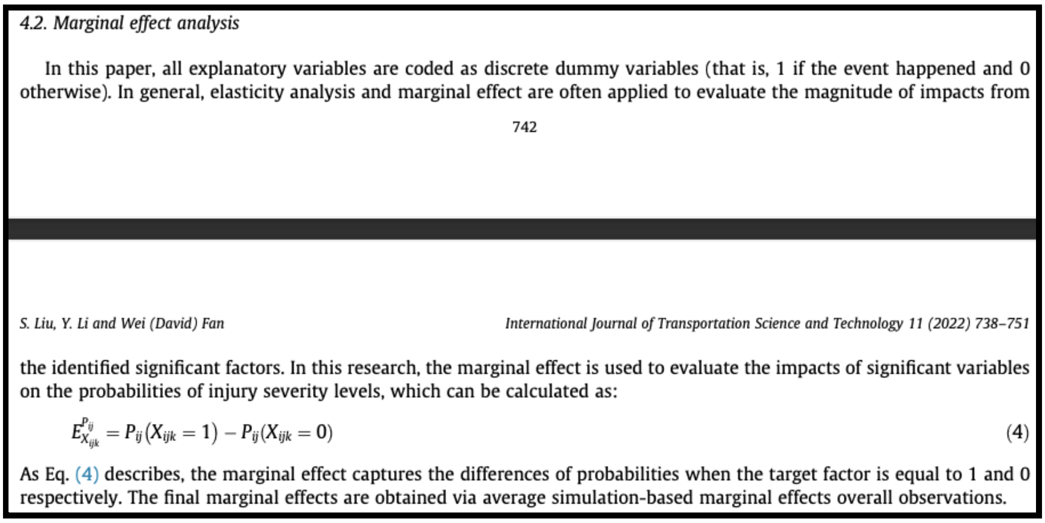
With the abovementioned setting, the probability of cyclist i sustaining injury severity j can be computed as:

$$P_{ij}|\beta_i = \frac{exp(\beta_i X_{ij})}{\sum_{j\in J} exp(\beta_i X_{ij})} \quad (2)$$

In the multinomial logit model, β_i is assumed to be fixed across individuals, which might not be valid considering the variation of sensitivities of individuals towards certain factors. For example, a driver with different driving skill levels may experience different challenges when he/she is driving in mountainous areas. Mixed logit model allows β_i vary across individual i by assuming the parameters following certain distribution. Then by considering the randomly distributed parameters across individual observations, a mixing distribution can be further written in Equation (3):

$$P_{ij} = \int (P_{ij}|\beta_i) f(\beta|\varphi) d\beta \quad (3)$$

Fig. 1.3. An extract taken from research paper regarding the mixed logit methodology used within the study [3]



4.2. Marginal effect analysis

In this paper, all explanatory variables are coded as discrete dummy variables (that is, 1 if the event happened and 0 otherwise). In general, elasticity analysis and marginal effect are often applied to evaluate the magnitude of impacts from

742

S. Liu, Y. Li and Wei (David) Fan — *International Journal of Transportation Science and Technology 11 (2022) 738–751*

the identified significant factors. In this research, the marginal effect is used to evaluate the impacts of significant variables on the probabilities of injury severity levels, which can be calculated as:

$$E^{P_{ij}}_{X_{ijk}} = P_{ij}(X_{ijk} = 1) - P_{ij}(X_{ijk} = 0) \quad (4)$$

As Eq. (4) describes, the marginal effect captures the differences of probabilities when the target factor is equal to 1 and 0 respectively. The final marginal effects are obtained via average simulation-based marginal effects overall observations.

Fig. 1.3. An extract taken from research paper regarding the marginal effect analysis methodology used within the study [4]

Pro requirement 2.2: Specifics regarding the software that are used within this study are not explicitly mentioned.

Presentation:

Pro requirement 1.9: The data analytics project must result in the “variables ~~are~~ identified with significant impacts on at least one of the cyclist injury severity levels”, “variables ~~(that is, male cyclist, cyclist on crosswalk, rural area, adverse road condition, and no traffic control) are~~ found to have random effects across all observations under different severity levels”, and “random parameter ~~has been~~ found in the nighttime model”.

Pro requirement 3.1: The causes that relate to the “Human characteristics”, “Vehicle characteristics”, “Environmental characteristics”, and “Crash characteristics” must be clearly defined for the “crashes involving cyclists in the daytime and nighttime separately”.
Gen requirement 3.1: The format by which the different categorical causes of the issue must be clearly stated when presenting the results of the analytics project.

Pro requirement 1.10: The data analytics project must “show the marginal effects for each significant contributing factor to the fatal injury of the cyclist in both models” using a histogram.

Case study 4: Diagnostic analysis of distributed input and parameter datasets in Mediterranean basin streamflow modeling [4]

Initiation:

Pro requirement 1.1: The goal of this analytics project is to “analyze the impact of different data sources in the input and parameterization phase of a water balance hydrological modeling application”.

Pro requirement 1.2: A table containing the nomenclature that relates to this data analytics project must be defined.

Pro requirement 1.3: This analytics project will analyze “different data sources in the input and parameterization phase of a water balance hydrological modeling application” as well “the comparison of different configurations of input and parameter datasets” .

Pro requirement 1.4: The data analytics project will use “LAI, reference evapo- transpiration, crop coefficients and volumetric soil moisture contents at wilting point, field capacity and saturation”.

Pro requirement 4.1: The results of the data analytics must be validated by applying the "model" "at daily scale in a semi-arid basin of Southern Italy (Carapelle river, basin area: 506 km2)".
Gen requirement 4.1: The analytics project must define how the results of the data analytics will be validated or verified.

Acquisition:
Pro requirement 2.1: For the evaluation of data sources "semi-distributed model was used, based on a discrete grid for the representation of vertical water fluxes (rainfall, evapotranspiration, infiltration and groundwater recharge) and a lumped representation of sub-horizontal fluxes (overland runoff, lateral flow and groundwater flow) "

Pro requirement 1.5: This analytics project will use: "Detailed data of watershed physical information, land uses and climate" taken from "Time series of rainfall, temperature and wind speed, recorded by the Hydrometric Office of Regione Puglia"; "Continuous streamflow data are provided by the gauging station at the Ordona Castelluccio dei Sauri bridge "; "Land use and vegetal coverage were obtained by the Corine land Cover (scale: 1:100,000)".

Project requirement 4.2: The derived data used within this analytics project is: "other climatic quantities required by the FAO Penman–Monteith equation were derived by temperature and wind speed"; "topographic features were defined using the Digital elevation map (90 m X 90 m) of the Carapelle watershed"; "Soil parameters such as the textural classes, saturated hydraulic conductivity, soil depths and porosity were extracted from the ACLA2 project "; "percentage of organic matter was derived from the Octop Project of the Euro- pean Soil Data Centre ".
Gen requirement 4.2: The derived data used in the analytics project must be defined.

Pro requirement 1.6: The "Main characteristics of the Carapelle watershed at Ordona Bridge closing section." is defined but the properties of the other data sets are not mentioned.

Analysis:
Pro requirement 1.7: The diagnostic analytics will be carried using "semi-distributed hydrologic model".

Pro requirement 1.8: The methodology has been defined but its complexity is beyond the scope of this thesis.

Gen requirement 2.2: The software or tool that were used within the analytics project are not explicitly mentioned.

Presentation:

Pro requirement 1.9: The data analytics project must result in the "evaluation of reference evapotranspiration ~~the FAO Pen- man–Monteith formulation~~ provided the best model performance", the best of "two different pedotransfer function sets provided" for "soil hydraulic properties", which of the "~~Among all~~ the metrics ~~the KGE~~ provided more sensitivity and convincing consistency with the recognized scientific value of the information and/or the methodology used to evaluate distributed model input and parameters."

Pro requirement 3.1: There is no well-defined formatting of the solution therefor the results are very hard to comprehend.

Pro requirement 1.10: The data analytics project must result in: graphs showing the "Duration curves using different values of the" of "subsurface flow coefficient c", "position parameter in the gamma distribution h", and "scale parameter kb in the gamma distribution"; as well as a "Comparison between observed and simulated discharges with the optimal dataset: hydrographs (a) and duration curves (b).".

Note: This case study was unique one given that it wasn't truly a diagnostic analytics project although the title claimed to be the diagnostic analysis of distributed input and parameter datasets, where as the result of this project was a diagnostic tool. Although for the purposes of this thesis this case study can be considered it does exceed the scope of this data analytics project. The scope of this case study shows the potential use of the generic requirements that are produced as a result of thesis to define what is required of a data analytics project with anything exceeding that being a beyond the scope of what a data analytics project is.

Case study 5: Diagnostic analysis of a single-cell Proton Exchange Membrane unitized regenerative fuel cell using numerical simulation [5]

Initiation:

Pro requirement 1.1: The goal of this project is to "to identify key performance limiting factors in fuel cell mode of a PEM Unitised Regenerative Fuel cell (URFC) fabricated at RMIT".

Pro requirement 1.2: A table containing the nomenclature that relates to this data analytics project must be defined alongside the “source terms of governing equations”.

Pro requirement 1.3: This data analytics project will carry out diagnostic analysis on the “single-cell PEM URFC designed and made at RMIT University”.

Pro requirement 1.4: The performance of the fuel cell is measured using the “maximum power of the RMIT cell ” in “W/cm2”.

Pro requirement 4.1: No validation methodology is mentioned.

Acquisition:

Pro requirement 2.1: The diagnostic analytics will be done using “computer simulatione namely the ANSYS PEM Fuel Cell Module”

Pro requirement 1.5: The data inputs for the model used within this project can only be acquired as derived data.

Pro requirement 4.2: The model used in this analytics project uses “estimated values of the input parameters obtained from” “the simulation polarization curve”

Pro requirement 1.6: The estimated data used in this diagnostic analytics project must contain data that relates to “model input parameters”.

Analysis:

Pro requirement 1.7: The diagnostic analytics project is carried out using the “diagnosis by simulation”.

Pro requirement 1.8: The steps for this project have been defined by Arif, Cheung et al. but the complexity of said steps are beyond the scope of this thesis.

Pro requirement 2.2: This diagnostic analytics project “ANSYS” to work with the “Fuel Cell Module”

Presentation:

Pro requirement 1.9: The results of the data analytics project include “the performance limiting factors of RMIT cell ”, the effects of “Hwang’s operating conditions ”, as well as varying “ the values of input parameters related to selected cell properties until the polarization curves of both URFCs matched in this region”.

Pro requirement 3.1: The results of the diagnostic project are not categorically defined.

Pro requirement 1.10: The diagnostic analytics project must line graphs that demonstrate the “Matching of simulated polarization curve RMIT single-cell PEM URFC with Hwang’s URFC experimental”.

2. Analysis of Case Study Analyses

Table 2.1 shows the requirements that have been elicited from each case study analysis. A total of 15 requirements have been elicited, with the first case study providing the highest number of generic requirements (10 requirements), and case study five being just used to validate the 15 requirements. The average number of requirements elicited within the first four case studies is 3.75, and the standard deviation is 4.19 whereas if you consider all five case studies then the mean and standard deviation becomes 3.00 and 4.00 respectively. The standard deviation being larger than the average meaning that there is a great variability in the number of new generic requirements that had to be defined in order to carry out the analytics project. The conclusions from these findings are incomplete without looking at the validation of the generic requirements.

It must also be noted that the complexity of requirements for diagnostic projects supersedes the complexity of those needed for descriptive analytics projects, shown by the mere amount of text required in order to define said requirements. This is a factor that needs to be taken into consideration when selecting requirement elicitation methods that are suitable for diagnostic analytics projects.

Table 2.1

Generic Requirements That Have Been Elicited From Each Case Study Analysis

Case study 1	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	1.10
Case study 2	2.1	2.2								
Case study 3	3.1									
Case study 4	4.1	4.2								
Case study 5	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a

Table 2.2 shows the results of the validation of generic requirements through the definition of project requirements.

Table 2.2

Validation of Generic Requirements That Have Been Elicited From Each Case Study Analysis

Gen (dia)	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	1.10	2.1	2.2	3.1	4.1	4.2
CS 1	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y					
CS 2	Y	Y	Y	Y	Y	N	Y	Y	Y	Y	Y	Y			
CS 3	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	N	Y		
CS 4	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	N	N	Y	Y
CS 5	Y	Y	Y	Y	N	Y	Y	Y	Y	Y	Y	Y	Y	N	Y

Legend: 🟨 Initiation Phase 🟩 Acquisition Phase

🟦 Analysis Phase 🟥 Presentation Phase

Y : The requirement was validated successfully

N : The requirement was not validated

The validation of Gen(dia) requirements shows that Gen(dia) requirements 1.1, 1.2, 1.3, 1.4, 1.5, 1.7, 1.8, 1.9, 1.10, 2.1, and 4.2 can be defined for most (80% of all case studies considered) if not all the analytics cases used for validation. Therefor it can be concluded that those generic requirements apply to all diagnostic analytics projects.

Gen(dia) requirement 1.6 is one that relates to the properties of the datasets used. The reason as to why this requirement could not be defined in some cases might be since the authors did not find it necessary to mention this information within the research article, but it is impossible to use a dataset without knowledge of the properties of the data set.

Gen(dia) requirement 2.1 is the only requirement that relates to the usage of models within diagnostic analytics projects. The models can be considered as a mix between a tool and a dataset in terms of its function within an analytics project but consideration of the need for acquiring the model from a specific source can be considered as part of the acquisition phase of the analytics project. The model also applies to the analysis phase since it influences the method selection process when carrying out the analysis.

Gen(dia) requirement 3.1 is one that relates to the categorical aspect of the solution that the diagnostic analytics project finds. Given that some analytics will result in the discovery that only one factor is affecting the performance whereas others will find multiple factors, Gen(dia) requirement 3.1 only applies in the latter case, but at the start of the project it is not possible to know the number of factors that the diagnostics will reveal. Therefore the recommendation can be made to consider this discussed with a stakeholder in advance to ensure that the opinions of the stakeholder and the analyst are aligned.

Gen(dia) requirement 4.1 is one that relates to the validation of the solution produced by the diagnostic analytics project. The need to validate the solution is something that stakeholders will decide based on the nature of the analytics project, therefore validating a solution does not need to be done unless explicitly mentioned by the clients since this can incur an additional cost of resources to the analytics project.

References

1. Fakhitah Ridzuan, Wan Mohd Nazmee Wan Zainon,Diagnostic analysis for outlier detection in big data analytics,Procedia Computer Science,Volume 197,2022,Pages 685-692,ISSN 1877-0509,

2. Lin Li, Fangjian Xie, Jingyi Li, Kangjia Gong, Xiaodong Xie, Yang Qin, Momei Qin, Jianlin Hu,Diagnostic analysis of regional ozone pollution in Yangtze River Delta, China: A case study in summer 2020,Science of The Total

3. Liu, S., Li, Y., Fan, W.D., *Mixed logit model based diagnostic analysis of bicycle-vehicle crashes at daytime and nighttime*, International Journal of Transportation Science and Technology, Volume 11, Issue 4, 2022, Pages 738-751, ISSN 2046-0430,https://doi.org/10.1016/j.ijtst.2021.10.001.

4. Milella, P., Bisantino, T., Gentile, F., Iacobellis, V. et al., *Diagnostic analysis of distributed input and parameter datasets in Mediterranean basin streamflow modeling*, Journal of Hydrology, Volumes 472–473, 2012, Pages 262-276, ISSN 0022-1694,https://doi.org/10.1016/j.jhydrol.2012.09.039.

5. Arif, M., Cheung, S.C.P., Andrews, J. , *Diagnostic analysis of a single-cell Proton Exchange Membrane unitized regenerative fuel cell using numerical simulation*, International Journal of Hydrogen Energy, Volume 46, Issue 57, 2021, Pages 29488-29500, ISSN 0360-3199, https://doi.org/10.1016/j.ijhydene.2020.11.165.