

# IMAGINING SOLUTIONS FOR EXTRACTING FURTHER VALUE FROM EXISTING DATASETS ON SURFACE AND GROUNDWATER RESOURCES IN SOUTHERN AFRICA

*Yannick Nuapia, Lindelwa Ndlhovu, Lungisa Ngundu, Phumlani Khoza, Anita Etale, Ewa Cukrowska and Hlanganani Tutu*



**USAID**  
FROM THE AMERICAN PEOPLE



science & innovation  
Department  
Science and Innovation  
REPUBLIC OF SOUTH AFRICA



**GROUNDWATER MANAGEMENT INSTITUTE**

**IBM Research | Africa**



**SWP**  
SUSTAINABLE WATER PARTNERSHIP

**USGS**  
science for a changing world



**WATER  
RESEARCH  
COMMISSION**

TT 842/20





# Imagining Solutions for Extracting Further Value from Existing Datasets on Surface and Groundwater Resources in Southern Africa

Report to the  
**Water Research Commission**

by

**Yannick Nuapia<sup>1</sup>, Lindelwa Ndlhovu<sup>1</sup>, Lungisa Ngundu<sup>1</sup>, Phumlani Khoza<sup>2</sup>,  
Anita Etale<sup>1</sup>, Ewa Cukrowska<sup>1</sup> and Hlanganani Tutu<sup>1</sup>**

<sup>1</sup>Molecular Sciences Institute, School of Chemistry, University of the Witwatersrand

<sup>2</sup>School of Computer Science and Applied Mathematics, University of the Witwatersrand

**WRC Report No. TT 842/20**

**ISBN 978-0-6392-0221-1**

**February 2021**



**USAID**  
FROM THE AMERICAN PEOPLE



science & innovation  
Department:  
Science and Innovation  
REPUBLIC OF SOUTH AFRICA



GROUNDWATER MANAGEMENT INSTITUTE

**IBM Research | Africa**



**Obtainable from**

Water Research Commission  
Private Bag X03  
Gezina  
Pretoria, 0031

[orders@wrc.org.za](mailto:orders@wrc.org.za) or download from [www.wrc.org.za](http://www.wrc.org.za)

This report forms part of a series of four reports. The other reports are:

- *Big Data Analytics and Modelling. Localising transboundary data sets in Southern Africa: A case study approach* (WRC Report no. TT 843/20)
- *Data Analytics and Transboundary Water Collaboration. Theme 1: Consolidation of Data and Application of Big Data Tools to Enhance National and Transboundary Data Sets in Southern Africa that Support Decision-Making for Security of Water Resources* (WRC Report no. TT 844/20)
- *Machine Learning Models for Groundwater Availability – Incorporating a Framework for a Sustainable Groundwater Strategy* (WRC Report no. TT 845/20)

**DISCLAIMER**

This report has been reviewed by the Water Research Commission (WRC) and approved for publication. Approval does not signify that the contents necessarily reflect the views and policies of the WRC, nor does mention of trade names or commercial products constitute endorsement or recommendation for use.

## EXECUTIVE SUMMARY

---

### BACKGROUND

Groundwater is an important source of freshwater for several semi-arid countries, including South Africa and Botswana. This is important for the rural communities that lie on the periphery of local water scheme pipelines and make direct use of groundwater in such regions. The Ramotswa Transboundary Aquifer (RTA) stretching between South Africa and Botswana supplies nearby rural communities, local municipalities and cities in both countries with fresh groundwater. Disparate water datasets have been compiled for the area, but with analytical tools such as big data analytics lacking that could yield valuable information for use in policy-making and water management. Further, other data sources such as citizen science do not feature which could help in deriving further value from datasets. To this end, the study (which formed Theme 2 in the broader programme) pursued the undermentioned aims.

### AIMS

- Engagement with stakeholders (water experts and municipal officers) to assess their understanding of the importance of big data analytics and citizen science.
- Collection of analytical water data; conducting a confirmatory survey; and collection of citizen science from the community.
- Conducting big data analytics on the data; text mining on citizen science; and building citizen science into the big data context.

To achieve these aims, the following approach to the research was used.

### METHODOLOGY

The approach involved conducting a survey of the perceptions of stakeholders (DWS; Rand Water, East Rand Water Care Company (ERWAT); academics in water and environmental research; and officials from the Ngaka Modiri Molema Municipality (NMMM) and the Ramotshere Moiloa Local Municipality (RMM)) to obtain some water data and other relevant information that was used to gauge their perceptions on big data and citizen science. In addition, a survey was done within the Wits School of Chemistry (for both students and staff) regarding their perceptions on the introduction of AI aspects to the chemistry curriculum. Citizen science data (i.e. perceptions regarding water) of the communities in the RMM (Zeerust area) and NMMM (Mahikeng area) municipalities were also solicited. Other stakeholders such as borehole sinking companies were also approached for more information about their operations and perceptions about water quality. The data from the different datasets (updated International Groundwater Assessment Centre (IGRAC) dataset, NMMM and 2 university theses) were used to obtain a consolidated dataset for which peculiar trends and patterns could be extracted. The dataset from NMMM (outside of the RTA) was used for the training as it was complete. Transfer learning (based on deep learning) was then conducted (using the R programming language) to predict missing values in the IGRAC dataset to which data from the theses had been added. This consolidated dataset

was then used for further data analytics using self-organising maps (SOMs) hybridised with k-means clustering.

Confirmatory sampling surveys were conducted for selected boreholes in both municipalities to substantiate some of the citizen science and analytical data. This water was collected at community boreholes located in community halls, clinics, schools and some private households. Field parameters (e.g. pH, electrical conductivity, temperature and total dissolved solids (TDS)), alkalinity, cation and anion concentrations were determined in the collected water samples. Hydrochemical models using the PHREEQC geochemical modelling code were conducted on the analytical results from the confirmatory samples. These included determining water hardness and usability for other purposes, e.g. agricultural (gardening). Some outcrop rock samples were also collected to assess their mineralogy. Confirmatory citizen science by the research team was conducted, that is, assessing the taste, colour, smell of the water and how it lathered with laundry soap. Other assessments, e.g. scaling on taps, tanks, toilet cisterns and kettles were made during the confirmatory surveys.

The data from IGRAC; data predicted from transfer learning; data from a confirmatory survey; and citizen science data were used towards text mining and modelling. Alkalinity values (for 872 samples) were used as the best tracker for the perceptions and comments collected in citizen science. Text was modelled using various text mining approaches and combinations, including establishing the corpus; data pre-processing and extracting the knowledge. Data pre-processing included transformation of text to numerical data and clustering aspects. The model accuracies were determined and refined for predictions of text classification. Similarity indices for rating comments were also determined which categorised the community responses into: negative (bad water quality), positive (good water quality) and neutral (indifferent about water quality).

It should be noted here that text mining is quite intricate, involving a number of steps that are inter-dependent. This would be expected as this type of analytics straddles unstructured and structured data and getting these to link to one another through a common platform.

## **RESULTS AND DISCUSSION**

There were mixed perceptions from the stakeholders regarding their understanding of big data (and artificial intelligence) and its importance in the water sector. Some were sceptical, citing the loss of jobs as their main concern with the technological revolution. Some felt that it was quite complicated to relate to and required mathematical skills to understand. There were others though that showed a keen interest, requesting training on these aspects. Some plans to train some of them on AI, machine learning and coding have been drafted. All the stakeholders in the water sector were supportive of the role that citizen science can play in complementing conventional analytical data. They expressed the view that where analytical data is scarce, citizen science could be used as a proxy. However, there were some that felt that citizen science should be verified as it could have biases. The survey showed that both staff and students from the chemistry background at universities were supportive of the introduction of a curriculum covering AI aspects. This has also been corroborated by increased numbers of students wanting to do short-term (vacation) projects with our research group on aspects of AI for the environment. They have alluded to uncertainty of jobs in future, with a likelihood

of increased demand of skills related to AI. Their colleagues in engineering, computer science and mathematics are already doing AI in their curricula and this has served to increase their interest as these are fields with high employment rates as well. To this end, the research team introduced an elective on AI module for chemistry in the BSc Honours curriculum and this is being offered in 2020.

The results from advanced data analytics showed that it was possible to identify different clusters in the water, using the groups: acceptable, borderline and unacceptable with respect to drinking water quality. Similar clustering was observed for the groupings with respect to agricultural use (mainly gardening). Some of the water, especially in the Mahikeng area showed high hardness. This corroborated some complaints that were gathered from some of the residents there regarding the unusability of the water for drinking purposes.

Transfer learning allowed for training of a complete dataset (from NMMM) and then using the parameters established there to make predictions for the missing cases for the IGRAC (mainly covering the RMM) dataset. The error of prediction was around 15%, making the predictions quite acceptable and the predicted data usable. The results from hydrochemical modelling showed that the majority of the water had acceptable quality. For water with unacceptable quality, high nitrate, calcium, chloride, sodium and bicarbonate concentrations were observed. The hardness was very high as well.

The majority of people (of a total of 60 interviewed) in the rural Zeerust area (Lekubu, Mokgole, Supingstad) were found to use water from community boreholes, e.g. boreholes located at community halls, clinics and schools. Their livestock drank from the puddles resulting from the water flowing from the borehole taps. A few households had their own private boreholes on their properties. These households were mostly for middle class or well-to-do families, e.g. teachers, school principals and people who were working in urban areas. The survey showed that people found the water usable, that is, from drinking to washing and gardening. The confirmatory survey also corroborated some of these responses, with most of the water sampled having a slightly salty taste (typical of borehole water in general), no smell and had clear colouration. However, the cumulative effects of scaling could be observed in toilet bowls and taps in one clinic and on some water storage tanks.

Similar surveys were made for the water in Zeerust town where they have a mixed source of water, namely borehole and surface water (both supplied by the municipality as piped water). This is treated through filtration and chlorination.

In Mahikeng, the water in Motsoseng and Magogoe townships (in both private and public boreholes) was found to be similar to that in rural Zeerust. It had a slightly salty taste, no smell and clear colour. The presence of dolomite ( $\text{CaMgCO}_3$ ) in these areas is quite apparent as the off-whitish rock outcrops can be seen on the surface everywhere. Analysis of the rock outcrop samples collected confirmed that it was dolomite, the main rock constituting the RTA. The water in Lotlhakane and some units in Mmabatho (e.g. Unit 1) was described by the residents as of bad quality. They indicated that they were not using it for drinking purposes, opting to purchase bottled water instead. One person was found to use commercial filters for their household drinking water in Mmabatho. It was clear from the descriptions that this was very hard water. This water was not sampled at the time of conducting the confirmatory survey and information regarding the water and boreholes in these areas could not be obtained from a borehole drilling company. As such, it remains a subject of further study.

It was possible to conduct text mining using the dataset constructed in the way described above. The relationship between textual and numerical data was established using alkalinity values which were in the ranges: 0-59 mg  $\text{l}^{-1}$ ; 60-120 mg  $\text{l}^{-1}$  and above 120 mg  $\text{l}^{-1}$ ). These were classified or rated as positive, neutral and negative, suggesting that the quality of the water was good, neutral and bad, respectively according to the perceptions. These, respectively, accounted for 58%, 32% and 10% of the samples in the dataset meaning that the majority of people found the quality of their water acceptable. This corroborated the findings of the confirmatory survey. The modelling accuracy for text mining was close to 100%, giving confidence that the predictions for text classification were quite accurate.

## CONCLUSIONS

Engagement with experts in the water sector showed that they were aware of AI and the potential of big data analytics in improving water management. Although some were sceptical about it, others showed a keen interest. They also had a general consensus that citizen science was important in complementing analytical water data, although citing the veracity of such data as one of the challenges to be taken into account. Citizen science indicated that most of the water in the area was of acceptable quality for household use. This was in agreement with confirmatory surveys that were conducted for the selected cases.

The patterns and trends in the clusters generated from the data analytics showed a general agreement with some of the observations from citizen science and confirmatory surveys. The majority of the water was clustered as of acceptable quality, which was corroborated by the citizen science in both rural and urban areas. There was no difference in water attributes observed between private and public boreholes.

Textual data (citizen science) was successfully linked to numerical data and these were modelled as a unit using text mining. It has been demonstrated that citizen science is important to consider when compiling datasets and conducting big data analytics. The good model metrics show that it can be used with confidence for unseen data or new cases. Stemming from this, further areas of collaboration with other stakeholders, e.g. related to the internet of things (IoT) in the water sector have been identified for future pursuit.

## RECOMMENDATIONS

Continued rapport with community members would help in dispelling the mistrust that they have about citizen science being potentially used for political purposes. Future work should investigate the extent of blending groundwater and surface water in urban areas and how this compares to the use of groundwater only. Further confirmatory surveys, especially in those isolated areas where the water quality was said to be very bad, should be conducted to understand why there are such discrepancies.

Text mining can be extended to make feeds such as social media (e.g. WhatsApp) useful as sources of information that can be screened the same way that was done for questionnaires. This is very common in marketing contexts. In the water context, communities can feed information about the quality of their water onto such platforms.

Another area that has potential for effective data collection is IoT where sensors can be used to monitor water quality in community tanks at depicted time intervals, with this data captured and modelled in real time. The

maintenance of these can be shared with communities, thus empowering them to participate in the management of their water.



## ACKNOWLEDGEMENTS

---

The project team wishes to thank the following people and organisations for their contributions to the project.

Reference Group	Affiliation
John Dini	Water Research Commission
Clara Bocchino	Sustainable Water Partnership
Shanna Nienaber	Water Research Commission
Sibusisiwe Makhanya	IBM Africa Research
Shafick Adams	Water Research Commission
Wandile Nomquphu	Water Research Commission
Department of Water and Sanitation; Department of Science and Innovation; USAID; International Union for Conservation of Nature (IUCN); SADC-GMI; IBM Africa Research	
Theme Teams	Affiliation
Theme 1	Umvoto Africa (Pty) Ltd
Theme 3	University of the Western Cape and L2K2 Consultants
Theme 4	Delta-H
Others	
Flora Makgale	Ngaka Modiri Molema Municipality (North West)
Difference Mokgalagadi	Ramotshere Moiloa Municipality (North West)
Obakeng Nchoe	University of the Witwatersrand (WITS)
Refilwe Setuki	University of the Witwatersrand (WITS)
Cornelius Rimayi	Department of Water and Sanitation
Sarah Ravhudzulo	Water Research Commission
Bonang Nkoane	University of Botswana
Ontibile Molefi	Environmental Consultant, Botswana
Shakera Arendze	Rand Water

The project was made possible by the collaboration of partners involved under the Big Data Analytics and Transboundary Water Collaboration for Southern Africa (refer to Chapter 1 for a detailed discussion).

# CONTENTS

---

<b>EXECUTIVE SUMMARY .....</b>	<b>i</b>
<b>ACKNOWLEDGEMENTS .....</b>	<b>vi</b>
<b>CONTENTS .....</b>	<b>vii</b>
<b>LIST OF TABLES.....</b>	<b>x</b>
<b>ACRONYMS &amp; ABBREVIATIONS .....</b>	<b>xi</b>
<b>GLOSSARY .....</b>	<b>xii</b>
<b>CHAPTER 1: INTRODUCTION AND BACKGROUND.....</b>	<b>1</b>
1.1 THE BIG DATA ANALYTICS AND TRANSBOUNDARY WATER COLLABORATION .....	1
1.1.1 Research projects: funding and training .....	3
1.2 PROJECT AIMS AND OBJECTIVES .....	4
1.3 PROJECT TEAM .....	4
1.3.1 University of the Witwatersrand .....	4
1.3.2 Collaborations .....	4
1.3.2.1 Municipalities .....	4
1.3.2.2 Department of Water and Sanitation .....	5
1.3.2.3 Rand Water.....	5
<b>CHAPTER 2: CASE STUDY OVERVIEW.....</b>	<b>6</b>
2.1 INTRODUCTION .....	6
<b>CHAPTER 3: OVERVIEW OF METHODOLOGY .....</b>	<b>10</b>
3.1 STAKEHOLDER ENGAGEMENT .....	10
3.1.1 Department of Water and Sanitation.....	10
3.1.2 Municipalities.....	10
3.1.3 Rand Water .....	10
3.1.4 Other stakeholders.....	10
3.2 WATER DATA COLLECTION .....	11
3.3 CITIZEN SCIENCE DATA COLLECTION .....	11
3.4 CONFIRMATORY SURVEY .....	12
<b>CHAPTER 4: BIG DATA AND DATA ANALYTICS.....</b>	<b>15</b>
4.1 WATER DATA FROM DATASETS.....	15
4.1.1 Extraction of data .....	15
4.1.2 Data pre-processing.....	16
4.1.3 Machine learning.....	16
4.1.3.1 Artificial neural networks.....	16
4.1.3.2 K-means clustering .....	17
4.1.3.3 Principal component analysis .....	18
4.2 DEEP LEARNING.....	18
4.2.1 Prediction of missing values using transfer learning.....	18
4.2.1.1 Performance evaluation.....	19
4.2.1.2 Performance evaluation.....	20
4.2.2 Self-organising maps .....	20
4.3 HYDROCHEMICAL ASSESSMENT.....	21

4.3.1	Water quality index .....	21
4.3.2	Irrigation water quality .....	22
4.4	CITIZEN SCIENCE AND DATA ANALYTICS.....	23
4.4.1	Activity 1: Establish the corpus .....	26
4.4.2	Activity 2: Pre-process the data .....	26
4.4.3	Activity 3: Extract the knowledge .....	29
<b>CHAPTER 5:</b>	<b>FINDINGS AND DISCUSSION .....</b>	<b>30</b>
5.1	PERCEPTIONS OF WATER EXPERTS .....	30
5.2	CITIZEN SCIENCE .....	33
5.3	CONFIRMATORY SURVEY .....	37
5.4	DATA ANALYTICS .....	38
5.4.1	Imputation of missing values.....	38
5.4.2	Self-organising maps for hydrochemical assessment .....	39
5.4.2.1	Assessment of groundwater quality for drinking purposes.....	40
5.4.2.2	Ion ratio coefficients.....	42
5.4.2.3	Suitability of the water for irrigation .....	42
5.4.3	Text mining for citizen science.....	43
5.4.3.1	Converting text to digital format.....	44
5.4.3.2	Cross validation .....	45
<b>CHAPTER 6:</b>	<b>REFLECTIONS ON LEARNING OPPORTUNITIES .....</b>	<b>48</b>
6.1	TRAINING PROGRAMMES .....	48
6.2	WORKSHOPS .....	48
6.3	CURRICULUM DEVELOPMENT .....	48
6.4	COMMUNITY INVOLVEMENT .....	49
6.5	FURTHER COLLABORATIONS.....	49
<b>CHAPTER 7:</b>	<b>CHALLENGES ENCOUNTERED.....</b>	<b>51</b>
7.1	STAKEHOLDER ENGAGEMENT .....	51
7.2	DATA-RELATED ASPECTS.....	51
<b>CHAPTER 8:</b>	<b>CONCLUSIONS .....</b>	<b>52</b>
<b>CHAPTER 9:</b>	<b>PROJECT RECOMMENDATIONS.....</b>	<b>54</b>
<b>REFERENCES</b>		<b>55</b>
<b>APPENDIX 1:</b>	<b>WATER EXPERTS SURVEY .....</b>	<b>60</b>
<b>APPENDIX 2:</b>	<b>CITIZEN SCIENCE DATA SURVEY .....</b>	<b>64</b>
<b>APPENDIX 3:</b>	<b>SUPPLEMENTARY DATA – R CODES USED.....</b>	<b>69</b>

## LIST OF FIGURES

---

Figure 2.1. Schematic example of an aquifer .....	7
Figure 2.2. Transboundary water basins between South Africa and neighbouring countries .....	7
Figure 3.1. Sites visited during confirmatory surveys.....	13
Figure 4.1. Structure of a multilayer perceptron .....	17
Figure 4.2. Voronoi cells in k-means clustering.....	17
Figure 4.3. Transfer learning process.....	20
Figure 4.4. Steps involved in text mining .....	23
Figure 4.5. A Venn diagram of the intersection of text mining and six related fields. ....	24
Figure 4.6. Text mining process .....	25
Figure 4.7. Activities involved in the context for text mining.....	26
Figure 4.8. Decomposition of pre-processing the data.....	27
Figure 5.1. Error minimisation during training of neural network.....	39
Figure 5.2. Clustering of borehole water using a self-organising map .....	40
Figure 5.3. k-means-SOM clustering for WQI, SAR, %Na and total hardness (TH). ....	41
Figure 5.4. Correlation matrix for the variable interrelationships.....	42
Figure 5.5. Cross validation plot .....	46
Figure 5.6. Distribution of comment rates using positive Cosine Similarity.....	47
Figure 6.1. Layout of the water quality monitoring system involving internet of things (IoT) .....	50



## LIST OF TABLES

---

Table 3.1. Summary of the variables from the dataset.....	11
Table 4.1. Relative weights of the variables in the study area .....	21
Table 4.2. WQI classification .....	22
Table 4.3. Classification of water for irrigation purposes.....	23
Table 5.1. Responses of water experts on big data analytics and artificial intelligence (AI).....	30
Table 5.2. Responses of the water experts on their perceptions of citizen science.....	31
Table 5.3. General assessment of community respondents' opinions.....	33
Table 5.4. Assessment of the community respondents' knowledge about their water .....	34
Table 5.5. Anion and cation levels in groundwater (nd – not detected; ppt – parts per thousand).....	37
Table 5.6. Confirmatory survey of alkalinity content in selected samples.....	38
Table 5.7. Accuracy results for different variables from modelling with Keras.....	39
Table 5.8. Text data obtained after converting numerical data .....	44
Table 5.9. Bag-of-words model .....	45
Table 5.10. Resampling results across tuning parameters .....	46

## ACRONYMS & ABBREVIATIONS

---

AI	Artificial intelligence
ANN	Artificial neural networks
DWS	Department of Water and Sanitation
EC	Electrical conductivity
FFNNs	Feed forward neural networks
IBM	International Business Machines
IE	Information extraction
IGRAC	International Groundwater Assessment Centre
IoT	Internet of things
IR	Information retrieval
IWMI	International Water Management Institute
LSA	Latent semantic analysis
ML	Machine learning
NLP	Natural language processing
NMMM	Ngaka Modiri Molema Municipality
PCA	Principal Component Analysis
ReLU	Rectified Linear Unit
RMM	Ramotshere Moiloa Municipality
RMSE	Root mean square error
RTA	Ramotswa Transboundary Aquifer
SOMs	Self-organising maps
SVD	Singular value decomposition
TDM	Total document matrix
TDS	Total dissolved solids
TH	Total hardness
WHO	World Health Organisation
WQI	Water Quality Index

## GLOSSARY

---

**Big data analytics:** is the use of advanced analytical techniques for very large, diverse data sets that include structured, semi-structured and unstructured data, from different sources, and of different sizes.

**Citizen science:** the collection and analysis of data by members of the general public. It is sometimes referred to as public participation in scientific research.

**Text mining:** also called text analytics, is an artificial intelligence (AI) technology that uses natural language processing (NLP) to transform the free (unstructured) text in documents and databases into normalised, structured data suitable for analysis by machine learning.

**Transfer learning:** is a research problem in machine learning (ML) that focuses on storing knowledge gained while solving one problem and applying it to a different but related problem.

## CHAPTER 1: INTRODUCTION AND BACKGROUND

---

### 1.1 THE BIG DATA ANALYTICS AND TRANSBOUNDARY WATER COLLABORATION

*This research project, managed by the Water Research Commission of South Africa, is part of a series of four projects under the Big Data Analytics and Transboundary Water Collaboration for Southern Africa, bringing together key stakeholders in Water and Big Data sectors.*

The *Collaboration* was first conceptualised in 2014 during the African Leaders Forum in Washington D.C., between USAID Global Development Lab and IBM Africa Research, which had opened its first hub in Nairobi (Kenya) in 2013, followed by the Johannesburg Lab in 2015. Since the early 2000s, the regional USAID mission for Southern Africa had been intensifying its regional support for transboundary water systems with both the Ramotswa Aquifer Project, involving Botswana and South Africa and the Resilience in the Limpopo River Basin Program (currently in its second phase with the Resilient Waters Programme, covering the entire Southern Africa region, with a focus on the Limpopo and Okavango River Systems). As part of this process, USAID had also been engaging with the Southern African Development Community (SADC) – Groundwater Management Institute and the Department of Science and Innovation of South Africa to support knowledge and technological advancement in the region. The focus of this multi-agency collaboration was agreed as Big Data Analytics and Transboundary Water. On April 3, 2017, the partners met with a multi-stakeholder regional group in a dynamic “Idea Jam” hosted by the IBM Africa Research Lab in Johannesburg. The objective was twofold:

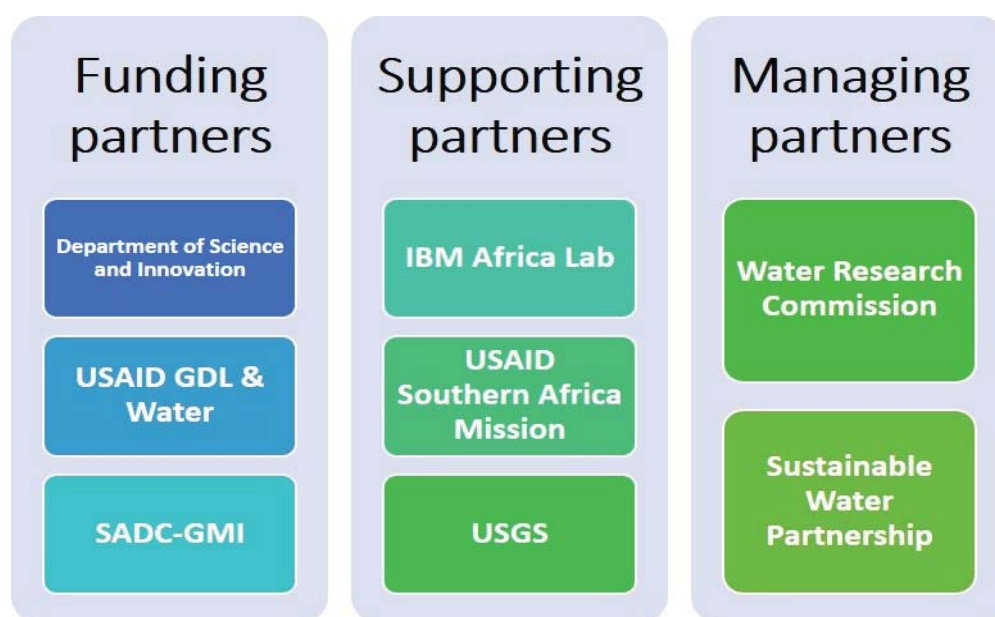
- to answer the broad question “how best can big data analytics be used to enhance transboundary water management”, and
- to identify the research questions, which would have guided the projects.

Requiring the collaboration of at least five high profile government agencies and private institutions, it took over one year to move from the Idea Jam to the launch of the Call for Proposals in August 2018, and the awarding of the four research projects in January 2019.

#### **The Collaboration: its partners and objectives**

Currently, the Collaboration has seven partners, with a joint function for USAID Global Development Lab, Water Office, and Southern African Mission. The partners each contributed to the development of the research projects based on own technical and funding capacity, see Figure 1.1. The total funds provided by the Funding Partners to research directly amount to USD \$500,000 (40%, 40%, 20%). IBM Africa contributed with the provision of their research facility in Johannesburg, *ad hoc*, but more importantly, by sponsoring the internship programme to the five candidates from the research projects.





**Figure 1.1. Collaboration partners and functions**

The Water Research Commission (WRC) was primarily tasked to oversee the financial and implementation management of the four research projects, as well as final reporting. The Sustainable Water Partnership (SWP) was called in by USAID in 2018 to act as the overarching Programme Coordinator, tasked with providing relation management, overall objective achievement, direction and positioning for the Collaboration in the region, and the fostering of a Community of Practice.

The United States Geological Service (USGS), IBM Research and SWP provided three sets of online training on issues pertaining to the focal topics of the Collaboration, which are now available on the Collaboration YouTube channel.

The Collaboration partners defined the objectives for this first phase of action (Table 1.1). However, the long-term vision is to create a Community of Practice for research and innovation on Big Data for Water Security, building on the multi-donor environment, which has proven successful.

**Table 1.1. Collaboration goals and objectives**

Goals	Objectives
Enhance current understanding of shared groundwater resources	Improve transboundary ground water management and collaboration
Provide big data skills development, capacity building and networking opportunities for Southern African researchers and their students	To foster multi-agency collaborative funding opportunities
To promote innovative thinking and application of Big Data Analytics to the Transboundary Water sector for integrated decision-making	To plant the seed for a growing community of pioneers in the use of Big Data Analytics for the study and management of Transboundary Water Aquifers

### 1.1.1 Research projects: funding and training

The four projects were awarded between December 2018 and January 2019, with a focus on a secondary river basin in the region: the Ramotswa, part of the Limpopo River Basin, spanning Botswana and South Africa. The lead institutions of the project teams have partnered (Figure 1.2) with the Botswana government and private institutions, as well as other leaders in previous water programme in the area, such as International Groundwater Resources Assessment Centre of the United Nations (UN-IGRAC, partner of Team 1) and International Water Management Institute (IWMI), implementers of the Ramotswa 2 USAID Project.

T1: Consolidation of data and application of big data tools to enhance national and transboundary data sets in Southern Africa that support decision-making for security of water resources.

- Umvoto Africa, University of Botswana, other global

T2: Consolidation of data and application of big data tools to enhance national and transboundary data sets in Southern Africa that support decision-making for security of water resources.

- Witwatersrand University, Geological Services of Botswana, DWS

T3: Localizing transboundary data sets in Southern African: A case study approach

- University of the Western Cape, CSIR, L2K2 Consultants

T4: Groundwater secure transboundary systems

- Delta-H Groundwater Systems and Institute for Groundwater Studies

**Figure 1.2. Titles of the four thematic areas and projects**

Despite working independently to address own project topics, the four research teams have progressively worked together to provide better integration for their outcomes. This process was led by the SWP in respect of providing a communication forum for the team leaders but was enhanced by the Internship Programme. The IBM mentors created a dedicated team and engaged the interns as individuals, as well as a group to help each other resolve new questions in coding and Machine Learning.

### Future prospects

As the current phase is coming to an end with the closing of the four research projects, the Collaboration partners are already identifying new opportunities to build on the lessons learnt and address the gaps recognised in this preliminary work, enhance the partnership to include national and regional government stakeholders, as well as new funding partners.

The focus of the Collaboration will remain the nexus between Big Data Analytics and (Transboundary) Water Security, recognising the inter-relatedness of successful water management in both national and shared aquifers to both human development and environmental goals.

## **1.2 PROJECT AIMS AND OBJECTIVES**

The overall aim of the project was to explore the potential of linking citizen science data with big data analytics as a way of extracting further value from conventional water datasets.

To achieve this, the following objectives were pursued:

1. To establish rapport with stakeholders (water experts, municipal officials, communities, academics, researchers and environmental consultants) to establish their understanding of big data analytics and citizen science and their relevance to the water sector.
2. To conduct big data analytics (based on machine learning and deep learning) on existing structured water datasets to predict missing values using transfer learning and to conduct data visualisation.
3. To conduct confirmatory surveys to establish a relationship between citizen science and analytical data.
4. To collect citizen science (unstructured) through interviews and integrate it into the big data framework using text mining.

## **1.3 PROJECT TEAM**

### **1.3.1 University of the Witwatersrand**

The project team consisted of: Profs Ewa Cukrowska, Luke Chimuka (both environmental chemists), Drs Anita Etale (expertise in community surveys and citizen science data collection), Yannick Nuapia (expertise in big data analytics), Phumlani Khoza (expertise in machine learning and programming) and Prof Hlanganani Tutu (team leader and expertise in environmental chemistry and big data analytics).

Students involved in the project were: Lindelwa Ndhlovu (BSc Hons) and Lungisa Ngundu (MSc and was sent to IBM Africa as an intern).

### **1.3.2 Collaborations**

Some collaborations that formed an important component of the project were established with relevant municipalities, the Department of Water and Sanitation (DWS), Rand Water and some individuals working for environmental consultancies and borehole drilling companies.

#### *1.3.2.1 Municipalities*

The following municipalities were collaborated with during the project: Ngaka Modiri Molema District Municipality (NMMDM), covering the greater Mahikeng area and the Ramotshere Moiloa Municipality (RMM), covering the Zeerust area. They provided data and some responses to questionnaires regarding big data analytics.

#### *1.3.2.2 Department of Water and Sanitation*

The DWS assisted with responses to some questionnaires on big data analytics and citizen science and provision of some of the water data. The project team also worked with one of their scientific technicians regarding possible workshops that the University could develop and offer on big data analytics with a focus on water management.

#### *1.3.2.3 Rand Water*

Rand Water provided responses to some questionnaires related to the application of big data analytics and citizen science to water management. The project team has since received a request to train staff there on big data analytics and coding. Further, there is a project that we have commenced for them on using historical data to improve water treatment plant performance.



## CHAPTER 2: CASE STUDY OVERVIEW

---

*This chapter presents an overview of the study area (the Ramotswa Transboundary Aquifer) with respect to water quality and challenges that have been faced thereof over the years. Several studies focusing on this aspect have been cited. The status of water quality data and if that has been harnessed for use in big data analytics is discussed. The use of citizen science or lack thereof has also been discussed.*

*Detailed descriptions and relevant literature related to the aquifer's geology, hydrogeology, groundwater levels and usage are available in Volume 1 of this four part series.*

### 2.1 INTRODUCTION

Groundwater is one of the most important water resources in several countries. It plays a crucial role in supplying water to people across the world with 25% of this population drinking water from karst groundwater resources (McGill et al., 2019; Bernard et al., 2015). In semi-arid countries, groundwater is a principal and an invaluable water resource. This is because of insufficient surface water availability due to unpredictable rainfall and excessive evapotranspiration rates.

Most of these groundwater resources or aquifers tend to cover large areas and as such contain large amounts of water. An aquifer is a permeable rock beneath the surface of the earth that is saturated with water (Figure 2.1). This water naturally comes to the surface via springs or can be pumped through boreholes drilled into the aquifer. The aquifers are recharged or replenished through rainfall and stream or river infiltration. In some cases, there may be artificial recharge through pumping of treated water, for instance, into the aquifer. Contaminants are usually transported with recharge water and the susceptibility of the aquifer depends on the aquifer material to transmit water. For instance, coarse materials such as gravel and sand tend to transmit water more rapidly than finer materials such as clay and silt.

The groundwater flow can be local, sub-regional or regional (i.e. transboundary). Transboundary aquifers (TBAs) are bodies of groundwater that cross into two or more international boundaries. In Africa, 72 TBAs have been mapped and cover 40% of the continent and supply water to a large portion of the population (Nijsten et al., 2018) in arid and semi-arid regions. Most of these TBAs are located in North Africa and the Sahel region. These water resources can be subject to conflicts of interest as a result of disproportionate resource partitioning as well as differences in social, economic and environmental management capacities of the sharing countries. On the contrary, TBA cooperation provides opportunities for country-to-country dialogue and sharing of data resources with the aim of improving the evaluation of the TBA and sustainable use of water (Braune and Christelis, 2014).

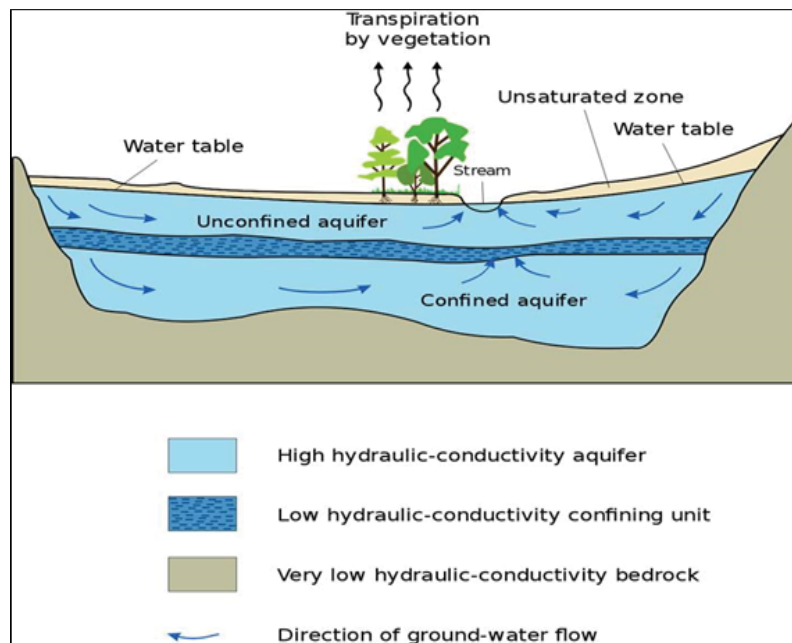


Figure 2.1. Schematic example of an aquifer (Alley et al., 1999)

A number of TBAs are found between South Africa and some of its neighbours (Figure 2.2). Most of these are located at water-stressed areas and are important sources of water for communities in those areas. One such TBA is the Ramotswa Transboundary Aquifer (RTA) that straddles South Africa and Botswana. This water resource has been important for either country in the recent years owing to population growth, industrial activities, agricultural activities, tourism and climate variability (e.g. prolonged drought spells). This has undoubtedly exerted pressure on the RTA.

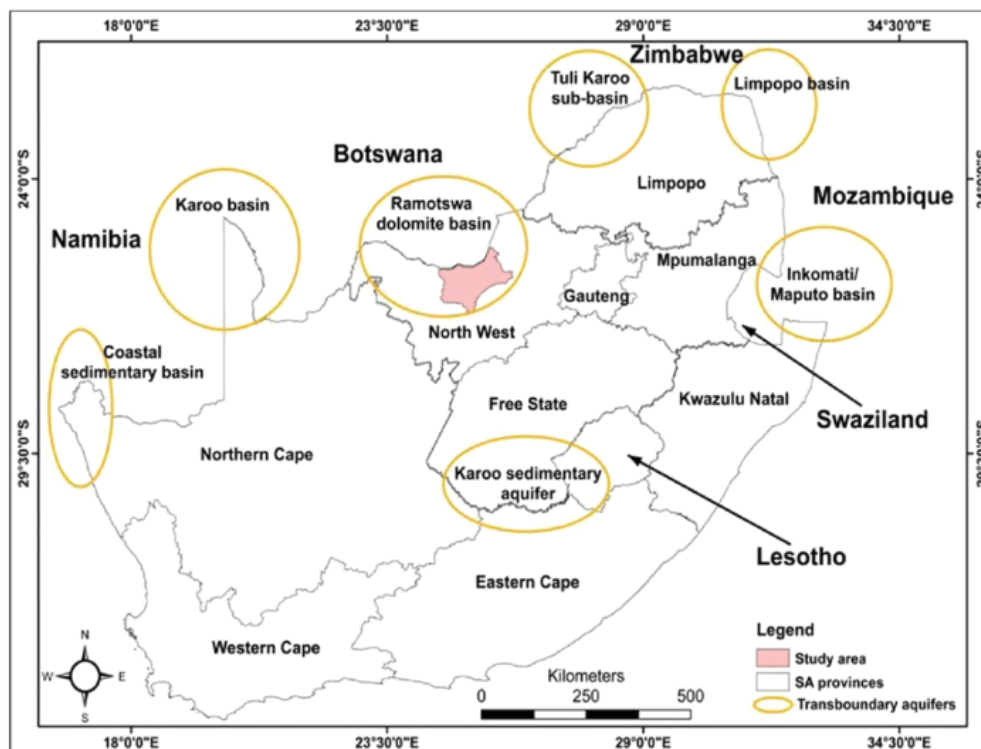


Figure 2.2. Transboundary water basins between South Africa and neighbouring countries

The RTA is a dolomitic (karst) aquifer and is one of the most important aquifers straddling South Africa and Botswana (Meyer, 2014; Pietersen et al., 2011). The aquifer exists within the upper parts of the Limpopo River Basin beneath the Marico and the Notwane subcatchments in South Africa and Botswana, respectively. The area lies within a relatively dry region in which rainfall is sporadic and unpredictably distributed with high seasonal variations between wet and dry seasons (Altchenko et al., 2016; Talma et al., 2013). Unequally distributed rainfall and elevated rates of evapotranspiration are the prevailing factors contributing to limited surface water resources (WRC, 2003). This has led to increased abstraction of groundwater by both Botswana and South Africa. In the former, abstraction had been curtailed in the 1990s owing to contamination of the aquifer by pit latrines and septic tanks (Owen, 2011, Weaver et al., 2007). However, abstraction resumed as a result of increased demand due to the population growth of Gaborone and other nearby towns as well as a need to supplement municipal water sources in the North West Province by South Africa (Villholth et al., 2013; Pietersen et al., 2011). The groundwater provides an ample supply throughout the year, supporting rural communities, small industries and agricultural activities (Ranganai et al., 2001; Villholth et al., 2013). Pollution from pit latrines in some areas and agricultural activities have posed a water quality problem for the aquifer, making it fail to meet requirements for drinking water in most cases (WHO, 2015). It was reported that approximately 3000 pit latrines constituted a major groundwater pollution hazard in and around the town of Ramotswa (Botswana) (Staudt, 2003). According to Davies et al. (2005), the use of pit latrines in human settlements within the aquifer area in both South Africa and Botswana was encouraged in the 1980s as a way of dealing with the problem of sanitation.

Some studies on water quality in the North West Province in general and within the aquifer specifically have pointed to nitrate ( $\text{NO}_3^-$ ) as the most common groundwater pollutant (BOS, 2000; Stadler et al., 2012; Wang et al., 2000; Vogel et al., 2004; Staudt and Anuraga et al., 2006). Stadler et al. (2012) attributed the increasing of nitrate concentration in the Ramotswa aquifer (South African side) to agricultural activities, influence from mines, animal waste, land use changes and sewage disposal. Nitrate ( $\text{NO}_3^-$ ) and ( $\text{NO}_2^-$ ) are the dominant and stable forms of nitrogen in water that contains dissolved oxygen and are easily distributed in groundwater with little or no retardation, that is, they are conservative ions (Weaver, 2007; Wang et al., 2000). Both forms can be lost from the groundwater system through the process of denitrification and, conversely, aerobic processes may result in the production of large quantities of nitrate and nitrite in groundwater or soil. Several other factors that influence the fate of nitrate and nitrite in the soil and groundwater include, but are not limited to: the depth of the water table, amount of rainfall, soil type and presence of organic material (Crafford et al., 2004; Wakida and Lerner, 2005; Anuraga et al., 2006). Nitrate has been recognised by many researchers as a common pollutant in groundwater around the world and has hazardous health consequences for both infants and adults when they are exposed to high concentrations (Knoll et al., 2019). The WHO guideline value for nitrate in drinking water is set at  $50 \text{ mg l}^{-1}$  as the majority of clinical cases of methemoglobinemia and hypertension being reported recorded nitrate levels greater than  $50 \text{ mg l}^{-1}$  (WHO, 2015).

Apart from the general pollution by nitrate, the groundwater is also exposed to bacterial contamination due to sanitation challenges that have been exacerbated by population growth and urbanisation (Wormald et al., 2003). Some studies have pointed to fairly high counts of total and faecal coliform in groundwater from some boreholes in Ditsobotla, Molopo and Zeerust in the North West Province (Kwenamore, 2006; Bezuidenhout

et al., 2011; Mulamattathil et al., 2000). The studies have cited human activities around the borehole areas as the main sources of contamination. Contamination of the aquifer by microbial and chemical constituents is quite apparent, although there are no studies that have been found that correlate land use patterns to the water quality. Also, there are no records of collection of citizen science data, that is, aesthetic and qualitative information of the water by the communities.

Understanding and assessing human activities in this transboundary aquifer area could provide a basis for improved decisions regarding the management of this precious groundwater resource for the benefit of both countries. To achieve this, a comprehensive approach is required to assess the data that has been compiled by a number of organisations such as the International Water Management Institute (IWMI); International Groundwater Assessment Centre (IGRAC); Water and Sanitation Departments in both South Africa and Botswana in studies conducted on water quality of the aquifer. The use of big data analytics tools and citizen science data can be explored in this regard.

Big data refers to large sets of complex data, both structured and unstructured, that traditional processing techniques and algorithms are unable to operate on. It is defined by volume, veracity, variation and velocity (Akbar et al., 2019). It is used in almost every department, including biology, social sciences and physics. A number of studies have reported the use of the big data analytics approach using machine learning techniques for water management and groundwater protection (Knoll et al., 2019; Akbar et al., 2019; Bernard et al., 2015). The results of these studies have helped in the development of policies for water management across boundaries and to avoid conflicts between water governing bodies of impacted countries. Such an approach has not yet been implemented for the RTA so as to establish a platform that can lead to improved policy formulation and better management and sharing of the water between South Africa and Botswana.

The pursuit of citizen science as a complementary aspect of big data and its use to glean further value from datasets has not been undertaken in the general literature and more specifically for the RTA. The growth of the area of deep learning in the big data analytics space provides potential, through text mining and image analysis, for incorporation of citizen science into the conventional data analytics tools.

Therefore, this study was focused on using available data from relevant datasets and applying big data analytics techniques to extract and understand the patterns within the data. Citizen science was collected and converted to forms that are compatible with those of conventional data and incorporated in the overall data analytics process.



## CHAPTER 3: OVERVIEW OF METHODOLOGY

---

*This chapter presents the approach(es) that was used in the study, namely: stakeholder engagement; water data collection and analysis; confirmatory survey; and citizen science collection and analysis.*

### 3.1 STAKEHOLDER ENGAGEMENT

The water data was mainly obtained from the IGRAC website which was updated by the Theme 1 team (complete dataset in Volume 1 of this series). For the purposes of addressing issues such as missing data values, extra data for aquifers similar to the RTA was sought from municipalities as well. The details of modelling of this data are provided at a later stage.

#### 3.1.1 Department of Water and Sanitation

Engagement with DWS was to gather their perceptions of water experts regarding artificial intelligence (AI), big data analytics and citizen science in water quality management. Four officials responded to a questionnaire related to this that was sent out to them.

#### 3.1.2 Municipalities

Some interviews were conducted with the municipal officials from NMMM (Mahikeng) and RMM (Zeerust) to assess their perceptions regarding AI, big data analytics and citizen science. Extra water data for other similar dolomitic aquifers within the region were obtained for use in conducting transfer learning with the aim to predict missing values for the RTA.

#### 3.1.3 Rand Water

Some interviews were conducted with scientific officers to assess their perceptions regarding AI, big data analytics and citizen science with respect to water management. The engagement led to workshops on coding and data analytics for water treatment being suggested so as to harness the power of these tools for better process management and decision making.

#### 3.1.4 Other stakeholders

Further engagements included with a borehole drilling company in the Mahikeng area to understand the water aspects from their perspective. Some environmental scientists from consulting companies were also engaged to understand their perceptions regarding AI, big data analytics and citizen science as well as general aspects about water that they encounter in their field of work.

A sample of the questionnaire used to collect information on perceptions and views related to AI, robotics and citizen science from the water experts is presented in Appendix 1.

### 3.2 WATER DATA COLLECTION

The datasets from IGRAC (expanded by the Theme 1 team), 2 university theses (University of the Witwatersrand) and NMMM municipality (for Mahikeng and Ditsobotla) were used. The data was combined into one dataset and then the missing cases removed to leave only available cases (a clean dataset). From the clean dataset, deep learning was used to train the cases and the results thereof used to predict the missing cases that had been removed. This will be explained in detail in the big data analytics section. The summary of the variables is presented in Table 3.1.

**Table 3.1. Summary of the variables from the dataset**

Variable number	Variable	Type	Range
1	Ca <sup>2+</sup>	Numeric	117 - 551.99
2	Mg <sup>2+</sup>	Numeric	7 - 241.58
3	K <sup>+</sup>	Numeric	0.100 - 134.35
4	Na <sup>+</sup>	Numeric	3.20 - 220.00
5	F <sup>-</sup>	Numeric	0.06 - 2.63
6	SO <sub>4</sub> <sup>2-</sup>	Numeric	0.300 - 163.00
7	Cl <sup>-</sup>	Numeric	0.37 - 372.00
8	CaCO <sub>3</sub>	Numeric	71.21 - 1619
9	EC	Numeric	7.50 - 321.45
10	TDS	Numeric	49.5 - 1166
11	NO <sub>3</sub> <sup>-</sup>	Numeric	0.02 - 123.66
12	pH	Numeric	6.44 - 716.00

### 3.3 CITIZEN SCIENCE DATA COLLECTION

Citizen science data was collected from communities in Zeerust (in town; Lekubu; Mokgola; Supingstad; Madikwe Game Reserve Gate; and Kopfontein border post). The areas where the data was collected include a lodge, tuckshops, taverns, a church, clinics, community halls, schools, public taps and private homes. The public areas here, e.g. clinics, community halls, taverns and schools are where most community boreholes are located. The people were interviewed while they came to collect water or to water their livestock and some were interviewed at their homes. The questions focused mainly on the aesthetic properties of the water, e.g. colour, smell, taste, lathering, scaling on the body and effects after drinking.

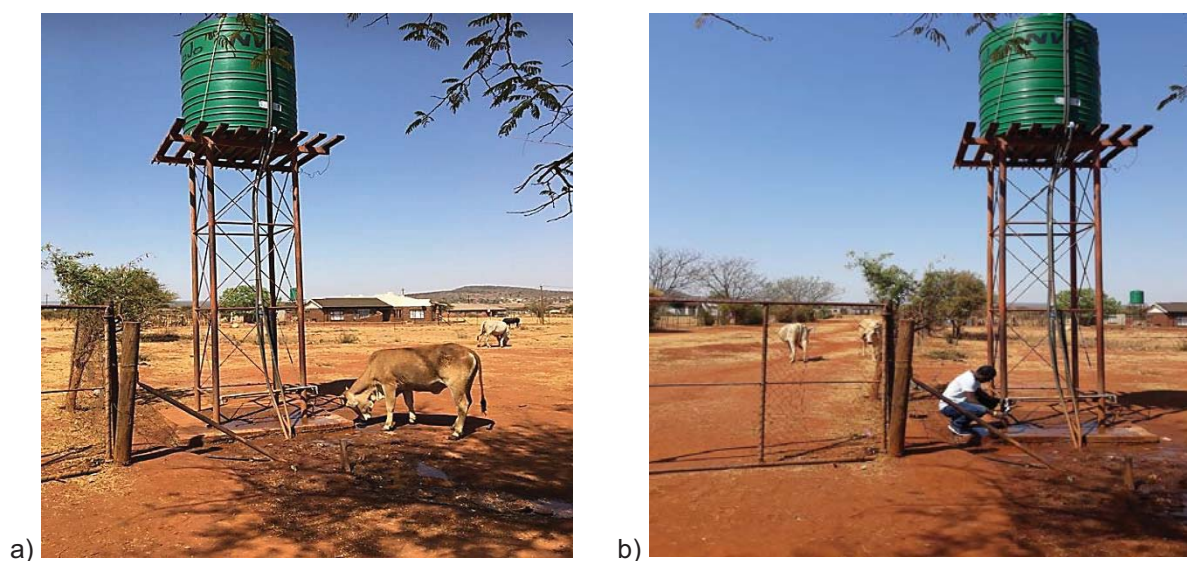
In Mahikeng, the citizen science was collected in town, and the townships of: Magogoe, Motsoseng, Lotlhakane, Montshiwa, Ramosadi and Goo-ra-makgetla. The areas included a lodge, schools, a tuck shop, a tavern and private homes. A total of 60 people were interviewed consisting of 30 women and 30 men. No children were interviewed as the ethics clearance did not allow for this.

A sample of the questionnaire used to collect information on perceptions and views related to citizen science from community members is presented in Appendix 2.

### 3.4 CONFIRMATORY SURVEY

A confirmatory survey was conducted together with citizen science data collection. Essentially, our research team observed respondents while they collected water and then approached them to ask about their perceptions regarding the water that they are using. The team then collected some samples into sampling containers and also checked the aesthetic properties independently (taste, smell and colour). Parameters such as pH, total dissolved solids (TDS) and electrical conductivity were taken *in situ* using field meters (Hannah Instruments). The samples were collected into acid washed and rinsed 100 ml polypropylene containers and stored in a cooler box that had ice packs prior to transportation to our laboratory at the University of the Witwatersrand, Johannesburg. In the laboratory, the samples were filtered and analysed for cations (Ca, Mg, Na and Fe among others) and anions (chlorides, fluorides and sulphates). Alkalinity, determined as  $\text{CaCO}_3$ , was determined using a potentiometric titrator (Metrohm), with  $0.01 \text{ mol l}^{-1} \text{ H}_2\text{SO}_4$  as a titrant. For the rest of the cases, the alkalinity was inferred (as total hardness) from hydrochemical modelling using the PHREEQC geochemical modelling code. Further, the analytical results were assessed to determine the Water Quality Index (WQI) and water hardness. Detailed descriptions of this are presented later.

Some of the areas visited during the survey are shown in Fig. 3.1.







**Figure 3.1. Sites visited during confirmatory surveys: (a) a cow drinking at a borehole in Lekubu, Zeerust (b) sampling at the borehole (c) community borehole tap in Supingstad (d) scaling on the sides of a water holding tank in Supingstad, Zeerust (e) Klein Maricopoort Dam in Zeerust, with visible efflorescent salts along its capillary fringe (f) borehole at GJ Podile School in Mahikeng (g and h) dolomite outcrop in Magogoe, Mahikeng**

No sampling was done at the Klein Maricopoort Dam as the banks were too muddy to allow for access to the water. This dam supplies part of the water for the Zeerust town community and this supply is complemented by borehole water.

## CHAPTER 4: BIG DATA AND DATA ANALYTICS

---

*This chapter describes the data treatment aspects related to the data collected from water datasets, citizen science and confirmatory surveys. Some theoretical aspects are presented and explained in the context of this study.*

*Data collection from the water datasets was followed by conducting a cleanup and structuring of the data; identifying and imputing missing data; and extracting patterns and trends within the data using clustering and classification techniques. Data clusters tend to yield important latent information in the dataset, e.g. areas that are most impacted (or hotspots). Citizen science data was converted to numerical forms that could be assessed using big data tools such as text mining. Hydrochemical modelling was conducted to further assess the chemistry of water with the aim to determine values for important missing variables. Information from the models was also used to infer potential citizen science responses based on the water quality parameters.*

### 4.1 WATER DATA FROM DATASETS

#### 4.1.1 Extraction of data

In this study, the R programming language was used to analyse data through the following steps: clean-up of the data, visualisation, prediction of missing values and clustering chemically related boreholes. R programming has slowly, yet surely, grown into a major source of statistical analysis in both research and academia (Horgan, 2012). It offers a wide range of packages and functions to help manipulate the data to better the research at hand or find solutions to other questions that might be of interest, something a conventional software cannot do. All the calculations come in-built into R and all one must do is call a function to activate them (Hamerly and Elkan, 2003).

The data collected from the IGRAC RIMS database (refer to Volume 1 of this series) was converted from an XLS file into a CSV file using Excel. The data was loaded onto R using the code shown below:

`ws = read.csv(file.choose(), header = TRUE, sep = ";", na.strings=c("", "NA"))`. The data was saved into variable **ws**. The **read.csv** function extracts the file as a CSV from Excel and the **header = TRUE** function identifies the file as having pre-defined headings for each column. The **na.strings = c("", "NA")** identifies all missing values as NA.



#### 4.1.2 Data pre-processing

The first step in data analysis is called data cleaning, which is a process of transforming raw data into consistent data that can be analysed by removing errors, missing data, duplicated variables, etc.

#### 4.1.3 Machine learning

Some machine learning tools were used to model the water data. These included feed forward neural networks (FFNNs) with back propagation; k-means clustering; principal component analysis; and self-organising maps (SOMs). Their theoretical overview is presented below.

##### 4.1.3.1 Artificial neural networks

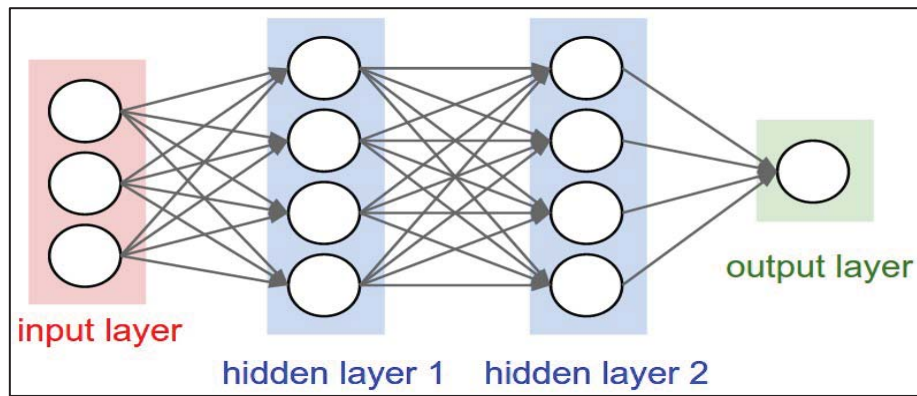
Artificial Neural Networks (ANN) were used for predictive analysis so as to impute the missing values. ANN is a machine learning technique used in computer systems for data analysis. In this study, multilayer perceptrons were employed (see Figure 4.1). This is a form of feedforward ANN where the connection between nodes is non-cyclic (Buscema et al., 2018). Learning occurs in the perceptron (nodes) by changing of weights until each piece of data (case) has been assessed. Error calculation then occurs when the difference between the target value and the resultant value is calculated. This is a form of supervised learning, where the machine learning algorithm tries to match a given input-output example from its own input-output calculations given by calculating the error as the difference:  $e_j = d_{nj} - y_{nj}$ . Thus, in ANN a complete set of data is used for training; and a set of data with missing values for testing. The training set is used to train the algorithm, and the testing set is to test how accurate it is. Minimization of error occurs after each case has been trialled and the error is given by:

$$\epsilon(n) = \frac{1}{2} \sum e_j^2(n) \quad (1)$$

Using gradient descent (first order algorithm for finding local minima), the change in each weight can be calculated as:

$$\Delta \text{weight}(n) = -\eta \frac{\partial \epsilon(n)}{\partial v_j} y_{nj}(n) \quad (2)$$

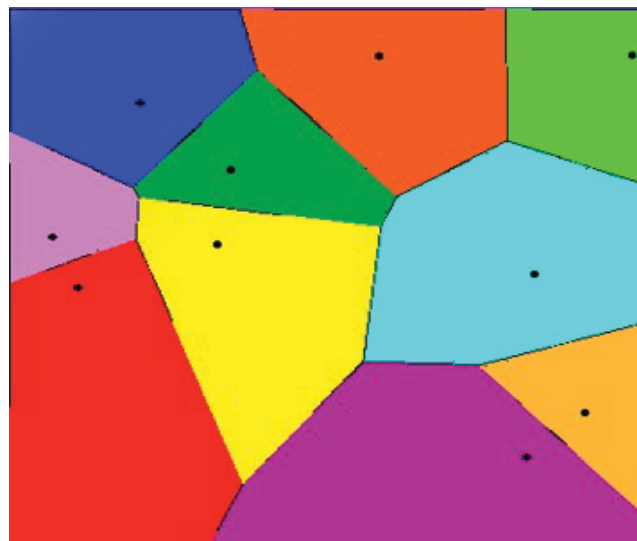
where  $v_j$  is the local field, which can vary and  $\eta$  is the learning rate (Buscema et al., 2018).



**Figure 4.1. Structure of a multilayer perceptron**

#### 4.1.3.2 K-means clustering

K-means clustering is a form of centroid clustering which aims to cluster  $n$  observations into  $k$  clusters with each observation belonging to the cluster with the nearest mean. This results in the data being partitioned into Voronoi cells (Figure 4.2) (Jain, 2010).



**Figure 4.2. Voronoi cells in k-means clustering**

The algorithm is as follows:

1. The value of  $k$  is randomly chosen.
2. Arithmetic  $k$  means are calculated to form randomly scattered centres of clusters.
3. The distance between the objects and centres are calculated using the Euclidean distance.
4. Each object is assigned to the nearest centre forming a cluster around it.
5. The cluster means are recalculated until the algorithm converges (the clusters remain the same).

#### 4.1.3.3 Principal component analysis

Principal Component Analysis (PCA) is also a clustering method slightly different from k-means. For in this method, the cases are clustered based on their new calculated scores (eigenvalues) and the variables are reduced into Principal Components.

For instance, PCA can take a collection of twenty variables and reduce them through dimension reduction to produce four Principal Components, which can account for all the variability in the data. The first principal component having the greatest variability is accounted for, then the second, the third and so forth (Smith, 2002). The following algorithm is used for PCA:

1. The data is converted into a matrix.
2. Calculate the mean of the matrix.
3. Calculate covariance matrix.
4. Calculate the eigenvectors and values.
5. Place the eigenvalues in descending order and use the ones that account for most data variability to form your PCs.
6. Transform the data into the new subspace of PCs.

## 4.2 DEEP LEARNING

### 4.2.1 Prediction of missing values using transfer learning

The data processing is summed up as follows.

The data consisted of 1326 boreholes with 17 variables (Borehole ID, Country, longitude, latitude,  $\text{Ca}^{2+}$ ,  $\text{Mg}^{2+}$ ,  $\text{K}^+$ ,  $\text{Na}^+$ ,  $\text{F}^-$ ,  $\text{SO}_4^{2-}$ ,  $\text{Cl}^-$ ,  $\text{CaCO}_3$ , EC, TDS,  $\text{NO}_3^-$ , and pH). However, 829 boreholes did not have data, 388 boreholes had all the complete variables and 109 boreholes had missing values. The boreholes with no data were removed from the data set and leaving behind a dataset of 497 boreholes. The data from 388 boreholes were normalised, split into training (50%), testing (30%) and validation (20%) datasets.

The following text briefly explains an important aspect of deep learning, namely Keras.

The Keras deep learning architectures (using the R coding language) were performed through the validation dataset while the performance was verified using the testing dataset. Since eight variables were missing from the 109 boreholes as described previously, 50 cases were randomly extracted from the testing dataset, and eight variables were removed. The resulting dataset, with missing values, was applied to assess the performance of the missing data approximation. The estimation of missing values was done variable by variable and not all at once.

The initial building block of Keras is a model, and the simplest model is called sequential. A sequential Keras model is a linear pipeline (a stack) of neural network layers. This code fragment defines a single layer with x artificial neurons, and it expects y input variables. Each neuron can be initialised with specific weights. Keras provide a few choices, the most common of which are listed as follows:

- ✓ Random\_uniform: Weights are initialised to uniformly small random values in the range  $(-0.05, 0.05)$ . In other words, any value within the given interval is equally likely to be drawn.
- ✓ Random\_normal: Weights are initialised according to a Gaussian model, with a zero mean and small standard deviation of  $0.05$ .
- ✓ Zero: All weights are initialised to zero.

The sigmoid is not the only kind of smooth activation function used for neural networks. Recently, with the introduction of Keras, a very simple function called the rectified linear unit (ReLU) has become popular because it generates very good experimental results.

The input layer has a neuron associated with individual observations. Typically, the values associated with each neuron are normalised in the range  $[0, 1]$  (which means that the dataset was scaled). The output layer is a single neuron with the activation function, Softmax, which is a generalisation of the sigmoid function. Softmax squashes a  $k$ -dimensional vector of arbitrary real values into a  $k$ -dimensional vector of real values in the range  $(0, 1)$ . Once defined, the model could be compiled so it could be executed by the Keras backend (either Theano or TensorFlow). During compilation the optimisation algorithm with a learning rate of  $0.001$  was used to update weights while the model was trained. The objective function was selected by using the optimiser to navigate the space of weights (frequently, objective functions are called loss function, and the process of optimisation is defined as a process of loss minimisation). The trained model was evaluated by calculating the mean squared error between the predictions and the true values. The trained model was compiled with a fit () function. The final trained model was then evaluated with the test set that contains new unseen cases.

#### 4.2.1.1 Performance evaluation

The potency of the missing value approximation was assessed using the Root Mean Square Error (RMSE), correlation coefficient and the relative prediction accuracy (A). The RMSE between the actual and predicted values indicates the capability of the prediction. It is computed as follows:

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (actual\ value - predicted\ value)^2}{number\ of\ missing\ values}} \quad (3)$$

The correlation coefficient calculates the linear similarity between actual and predicted values. It ranges from  $-1$  to  $1$ , where its absolute value relies on the strength of the correlation. A value close to  $1$  shows a reliable predictive capability (Lin et al., 2018). The formula is given by (6), where  $x$  is the mean of the data.

$$r = \frac{\sum_{i=1}^n (actual\ value - mean\ values)(predicted\ value - means)}{[\sum_{i=1}^n (actual\ value - mean\ value)^2 \sum_{i=1}^n (predicted\ value - mean\ value)^2]^{\frac{1}{2}}} \quad (4)$$

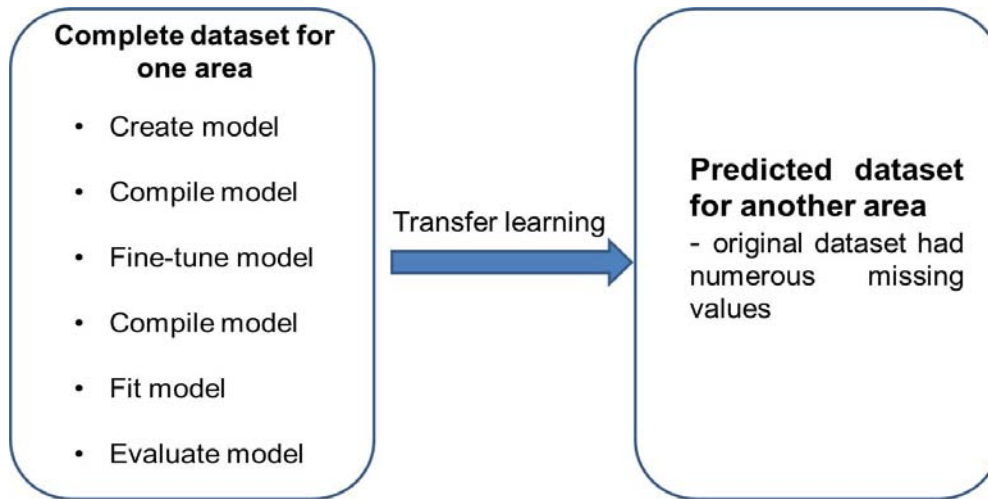
The relative prediction accuracy measures the number of predictions made within a certain tolerance. The tolerance is based on the sensitivity needed by the application. For the current dataset, the tolerance was set to  $10\%$ , as reported by Shukur and Lee (2015). The accuracy is computed as follows:

$$A = \frac{\text{number of predictions within tolerance} \times 100}{\text{number of missing values}} \quad (5)$$

Applying the performance parameter as mentioned above, the approximation of missing values was assessed by estimating each of the eight attributes individually.

#### 4.2.1.2 Performance evaluation

The process of transfer learning can be summarised as shown in Fig. 4.3. In the study, the training was conducted on a complete dataset from the NMMM (with most cases outside the aquifer area). The geology in which the cases are located is similar to that for the aquifer area, predominantly dolomitic. The trained model was then transferred to make predictions for the numerous missing cases in the original IGRAC database (the data incompleteness of this database was indicated in previous deliverable). Then the complete dataset following transfer learning was used for further data analytics.



**Figure 4.3. Transfer learning process**

#### 4.2.2 Self-organising maps

Although these would normally be grouped under ML techniques, they have been included here as a result of their hybridised nature in this case. To improve the mapping of the groundwater samples, a hybrid k-means-SOM was modelled in R. SOMs are common under the machine learning techniques that use unsupervised learning. In this case, the inputs are mapped directly into outputs by assigning them labels. The concept used is that of “neighbourhood”. In other words, an input that depicts say a high concentration of a variable is assigned a high weight and is considered a “winner”. The other inputs are then clustered around it by order of their weights, i.e. those with elevated concentrations will be very close to the winner while those with lower concentrations will be located farther away. This way, data clusters can be created that show the data visualisation a lot clearer. K-means is a similar technique but uses data centroids as described in the previous report. The combination of these two methods allows for refinement of the clustering, thus enhancing the models. Such hybridisation is very common in data analytics as it combines the capabilities of different techniques, giving better models than the individual techniques.

### 4.3 HYDROCHEMICAL ASSESSMENT

#### 4.3.1 Water quality index

The quality of groundwater is essential since it indicates the suitability of water used for irrigation and drinking purposes (Zheng et al., 2017). The water quality index (WQI) is an important parameter to assess groundwater quality and its relevance for drinking purposes. This index ranged from 0 (poor) to 100 (ideal), based on water quality variables weighted according to their relative importance. To compute the WQI, four steps were followed. In the first step, each of the parameters (pH, EC, TDS,  $\text{Ca}^{2+}$ ,  $\text{Mg}^{2+}$ ,  $\text{Na}^+$ ,  $\text{K}^+$ ,  $\text{Cl}^-$ ,  $\text{SO}_4^{2-}$  and  $\text{NO}_3^-$ ) was assigned a weight ( $w_i$ ) according to its relative contribution in the overall quality of groundwater for drinking purposes (Table 4.1). In the second step, the relative weight ( $W_i$ ) of each parameter is computed from the following equation:

$$W_i = \frac{w_i}{\sum_{i=1}^n w_i} \quad (6)$$

where  $w_i$  is the weight of each parameter, and  $n$  is the number of parameters.

**Table 4.1. Relative weights of the variables in the study area**

Variable	SI	Weight	Relative weight
$\text{Ca}^{2+}$	75	5	0.116
$\text{Mg}^{2+}$	50	4	0.093
$\text{K}^+$	12	2	0.046
$\text{Na}^+$	200	2	0.046
$\text{F}^-$	1	5	0.116
$\text{SO}_4^{2-}$	250	3	0.069
$\text{Cl}^-$	250	5	0.116
$\text{CaCO}_3$	300	2	0.046
EC	500	5	0.116
TDS	500	5	0.116
$\text{NO}_3^-$	45	2	0.046
pH	8.5	3	0.069

In the third stage, the quality rating scale ( $q_i$ ) for each parameter is allocated by dividing its concentration in each water sample by its relevant standard according to the guidelines laid down in the WHO (2004), and the result is multiplied by 100:

$$q_i = \frac{C_i}{S_i} \times 100 \quad (7)$$



where  $C_i$  is the concentration of each parameter in each groundwater sample in  $\text{mg l}^{-1}$ , and  $S_i$  is the WHO standard for each chemical parameter. In the last step, the WQI is evaluated as follows:

$$WQI = \sum_{i=1}^n W_i \times q_i \quad (8)$$

WQI classifies the water samples within five classes as described in Table 4.2.

**Table 4.2. WQI classification**

Range	Type of water
<50	Excellent water
50 - 100	Good water
100 - 200	Poor water
200 - 300	Very poor water
>300	Water unsuitable for drinking purposes

#### 4.3.2 Irrigation water quality

The suitability of groundwater for agricultural purposes depends on the effect of mineral constituents of water on both plants and soil. Effects of salts on soils causing changes in its structure, permeability and aeration indirectly affect plant growth. There is a significant correlation between sodium adsorption ratio (SAR) values for irrigation water and the extent to which sodium is adsorbed by the soils. If the water used for irrigation has a high concentration of sodium and low level of calcium, the cation exchange complex can become saturated with sodium, which can destroy the soil structure owing to the dispersion of clay particles (Belkhiri and Mouni, 2012). SAR was computed as follows:

$$SAR = \frac{Na^+}{\sqrt{\frac{(Ca^{2+} + Mg^{2+})}{2}}} \quad (9)$$

SAR values less than 10 indicate excellent water quality; 10-18: good water; and >18 doubtful to unsuitable for irrigation purposes. A high concentration of SAR leads to the development of an alkaline soil, which will be hard and compact when dry and increasingly impervious to water penetration.

The concentration of the sodium in irrigation water is known as the sodium percentage (%Na). The sodium percentage has been used to classify the chemical composition of the groundwater. Excess sodium in water will change the soil structure and reduce soil permeability. The sodium percentage can be calculated as follows:

$$Na\% = \frac{(Na^+ + K^+) \times 100}{Ca^{2+} + Mg^{2+} + Na^+ + K^+} \quad (10)$$

The classification of groundwater, according to their %Na is presented in Table 4.3.

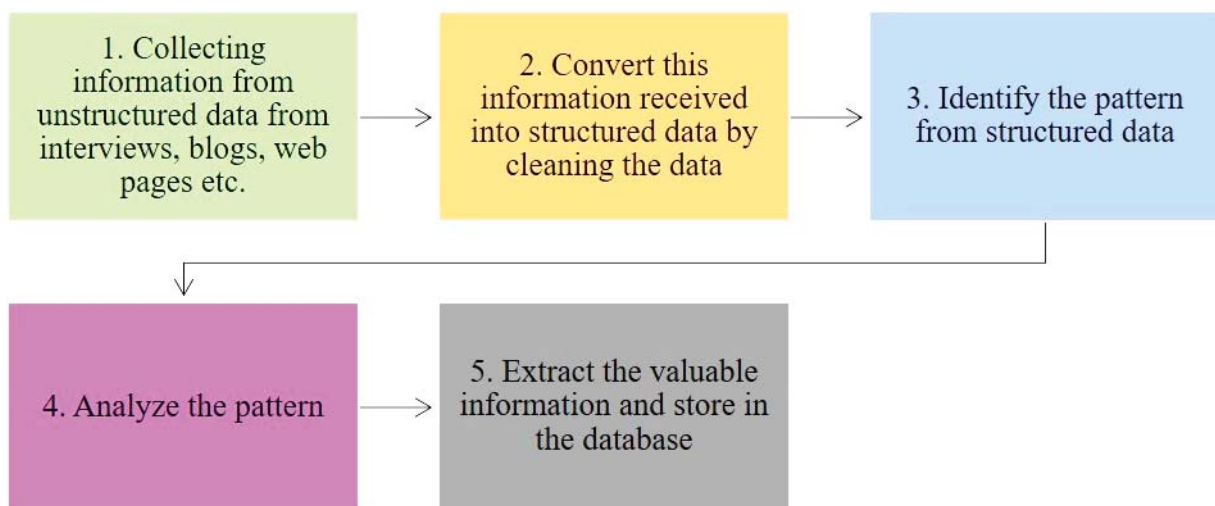
**Table 4.3. Classification of water for irrigation purposes**

%Na	Water class
<20	Excellent
20 - 40	Good
40 - 60	Permissible
60 - 80	Doubtful
>80	Water unsuitable for irrigation

#### 4.4 CITIZEN SCIENCE AND DATA ANALYTICS

Citizen science (from questionnaires to communities) and analytical water data (from confirmatory surveys and datasets completed through transfer learning) were used for text mining. The following discusses text mining and the general methodology involved, which was adopted for this study.

Text mining (also referred to as text analysis) is a process of extracting interesting and significant patterns or numeric indices by means of identifying facts, relationships and assertions within textual data (Soumen, 2002). Text mining makes text accessible to various algorithms (e.g. machine learning algorithms) for further analysis (Bruce et al., 2009). According to Manning et al. (2008), it has become a very important technique in the analytics industry due to the availability of unstructured text data from multiple sources (e.g. interviews, surveys, SMSs and product reviews, to name a few). It is a multi-disciplinary field based on information retrieval, data mining, statistics, and computational linguistics (Nisbet et al., 2009). Similar to other techniques in artificial intelligence (AI), text mining has fundamental steps that have to be followed (Figure 4.4).



**Figure 4.4. Steps involved in text mining**

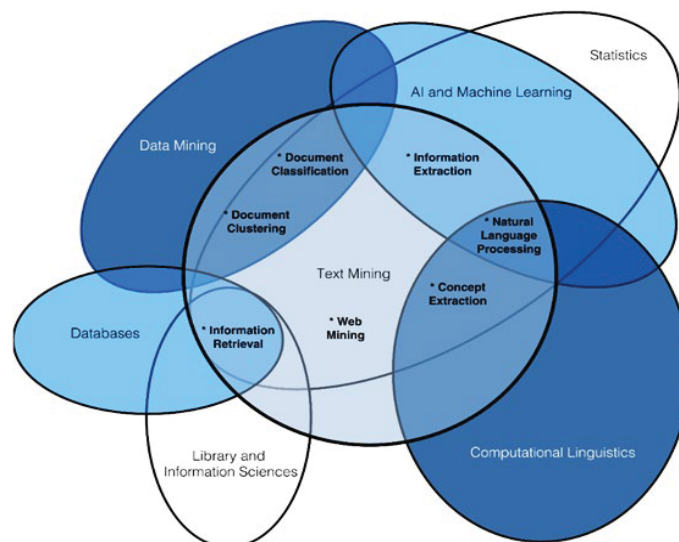
Text mining consists of five well-known techniques, one of which is Information Retrieval (IR). IR is a process of extracting relevant information from textual database based on the user's query (Smiley et al., 2009). The most renowned IR systems are Google and Yahoo search engines which recognize those documents on the

World Wide Web that are associated to a set of given words (Soumen, 2002). Information extraction (IE) is a method of extracting semantic (logic or context) information from textual data by identifying relationships and patterns within semi-structured or unstructured text. IR works best for the extraction of valuable information from structured data and IE for extraction from unstructured data (Nisbet et al., 2009).

Categorisation is a type of supervised learning where categories are known in advance and is essentially a process of gathering, processing and analyzing text documents to put them in the correct category based on the link between the content and category (e.g. spam filtering based on content) (Horto et al., 2003). Clustering is an unsupervised technique that finds intrinsic structures in information and arranges them into subgroups called clusters to generate labels for the objects based on the data (Lochbaum et al., 2000). Cluster analysis can be used as a standalone text mining tool to achieve data distribution, or as a pre-processing step for other text mining algorithms applied to the detected clusters. Text summarisation is the process of automatically creating a compressed version of a given text that provides useful information for the user (Ratinoy and Roth, 2009). A summary is a text that is produced from one or more texts that contains a significant portion of the information, reduced in length and keeps the overall meaning as it is in the original texts (Renders, 2004).

Text mining and analytics are umbrella terms describing a range of technologies for analysing and processing unstructured and semi-structured text data. The unifying theme behind each of these approaches is the need to turn text into a digital form that can be processed by algorithms. Converting text into a structured, numerical format and applying analytical algorithms requires knowledge on how to combine techniques for handling text, ranging from individual words to documents to entire document databases (Smith and Humphreys, 2006).

Currently, text mining has not had comprehensive definition because the field emerges out of a group of related but distinct (Vidhya and Aghila, 2010). Due to the breadth and disparity of the contributing disciplines, it can be difficult even to concisely characterize or define the term text mining. There are seven different text mining practice areas (Figure 4.5).



**Figure 4.5. A Venn diagram of the intersection of text mining and six related fields (shown as ovals), such as data mining, statistics, and computational linguistics. The seven text mining practice areas exist at the major intersections of text mining with its six related fields (Vidhya and Aghila, 2010).**

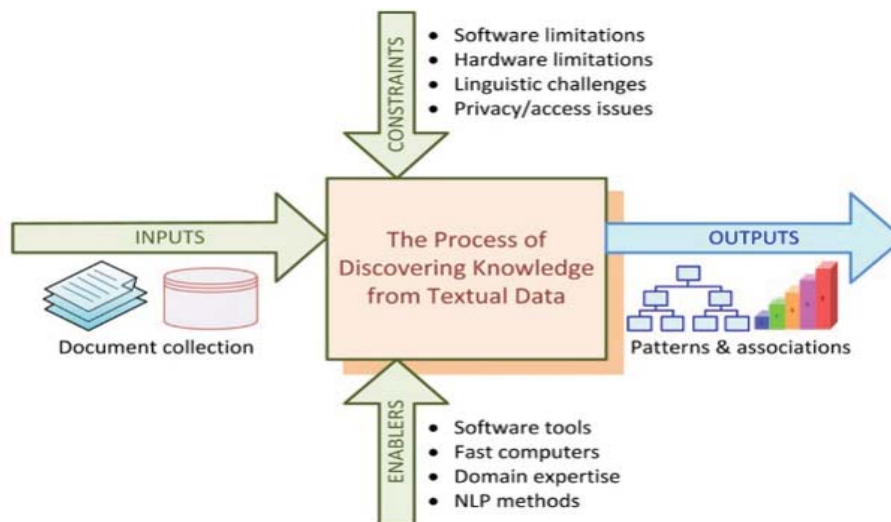
The primary purpose of text mining is to process unstructured (textual) data and structured and semi-structured data (if relevant to the problem being addressed) to extract novel, meaningful, and actionable knowledge/information for better decision-making (Dale et al., 2000).

One of the challenges of text mining is converting unstructured and semi-structured text into a structured vector-space model. This must be done prior to doing any advanced text mining or analytics (Diligenti et al., 2000). The possible steps of text pre-processing are the same for all text mining tasks, though the choice of the processing steps depends on the task. The basic steps are as follows:

1. Choose the scope of the text to be processed (documents, paragraphs)
2. Tokenise: Break text into discrete words called tokens.
3. Remove stop-words (“stopping”): Remove common words such as “the”.
4. Stem: Remove prefixes and suffixes to normalise words for example, run, running, and runs would all be stemmed to run.
5. Normalise spelling: Unify misspellings and other spelling variations into a single token.
6. Detect sentence boundaries: Mark the ends of sentences.

Normalise case: Convert the text to either all lower or all upper case.

The process involves inputs in the form of unstructured, semi-structured, or structured data collected, stored, and processed (Figure 4.6). The outputs represent the context-specific knowledge products that can be used for decision-making. The constraints (or controls) arrow entering at the top edge of the box represents software and hardware limitations, privacy issues, and the difficulties related to processing of the text that is presented in the form of natural language. The enablers entering the bottom of the box represent software tools, fast computers, domain expertise and natural language processing (NLP) methods.



**Figure 4.6. Text mining process (Sureka and Varma, 2008)**

The processes above (Figure 4.6) can be decomposed into three linked sub-processes that are called activities (Figure 4.7). Each has inputs, accomplishes some transformative process, and generates various outputs. If, for some reason, the output of a sub-process is not what was expected or emerges at an unsatisfactory level, feedback loops redirect information flow to a previous task to permit adjustments and corrections.

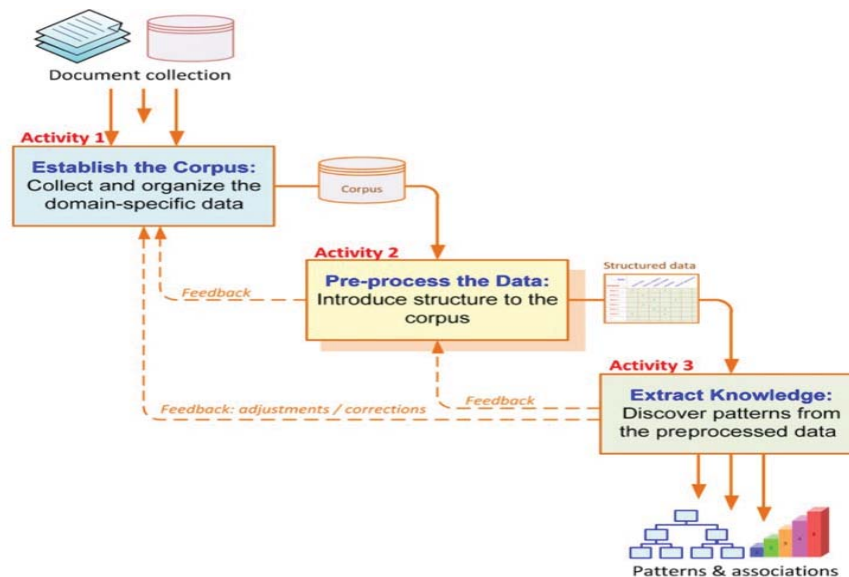


Figure 4.7. Activities involved in the context for text mining (Li et al., 2010)

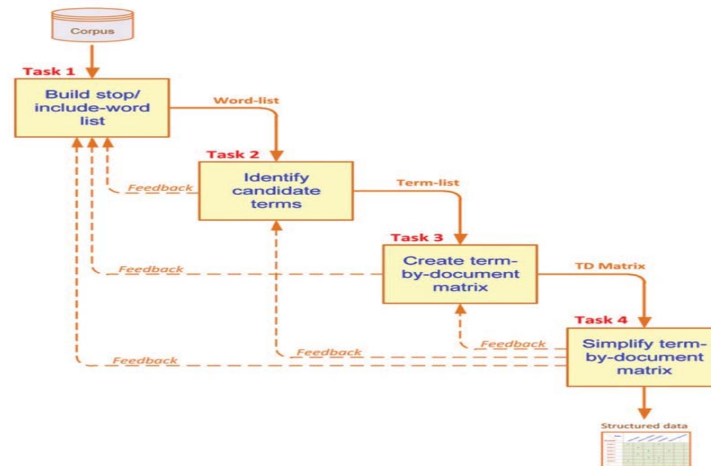
#### 4.4.1 Activity 1: Establish the corpus

The purpose of Activity 1 is to collect all of the documents that are relevant to the problem being addressed (see Figure 4.7). The quality and quantity of the data are the most important elements of both data mining and text mining projects (Seni and Elder, 2011; Polon, 2011). Sometimes in a text mining project, the document collection is readily available and is accompanied by the project description, for example conducting sentiment analysis on customer reviews of a specific product or service (Lohr, 2009). But, usually the text miner is required to identify and collect the problem-specific document using either manual (e.g. interviews) or automated techniques such as a web crawler that periodically collects relevant news excerpts from several websites. Data collection can include textual documents, HTML files, emails, web posts, and short notes (Hill and Lewicki, 2007). In addition to normal textual data, voice recordings may be included by transcribing using speech recognition algorithms. Once collected, the text documents are transformed and organized in a manner such that they are all represented in the same form, for instance ASCII text files for computer processing (Hu and Liu, 2004). The organization of these documents can be as simple as a collection of digitised text excerpts stored in a file folder, or it can be a list of links to a collection of web pages in a specific domain.

#### 4.4.2 Activity 2: Pre-process the data

In this activity, the digitised and organised documents (the corpus) are used to create a structured representation of the data, often referred to as the term-document matrix (TDM). The TDM consists of rows represented by documents and columns representing terms (Hill and Lewicki, 2007). The relationships between the terms and the documents are characterised by indices, which are relational measures, such as how frequently a given term occurs in a document (Seni and Elder, 2011). The goal of Activity 2 is to convert the list of organized documents (the corpus) into a TDM where the cells are filled with the most appropriate indices. The assumption made here is that the meaning of a document can be represented with a list and frequency of the terms used in that document. But, not all terms are equally important when characterizing

documents. Some terms, such as articles, auxiliary verbs, and terms used in almost all of the documents in the corpus, have no distinguishing power and therefore should be excluded from the indexing process (Sureka and Varma, 2008). This list of terms, commonly called stopwords, is often specific to the domain of study and should be identified by the domain experts. On the other hand, one might choose a set of predetermined terms under which the documents are to be indexed; this list of terms is conveniently called include terms or dictionary. Additionally, synonyms or pairs of terms that are to be treated the same and specific phrases can also be provided so the index entries are more accurate. A more detailed view of the TDM with its four tasks is presented (Figure 4.8) (Diligenti et al., 2000; Mahgoub et al., 2008).



**Figure 4.8. Decomposition of pre-processing the data (Hadley, 2011)**

#### Task 1

The first task generates stopwords along with synonyms and specific phrases.

#### Task 2

The term list is created by stemming or lemmatization, which refers to the reduction of terms to their simplest forms (i.e., roots). An example of stemming is to identify and index different grammatical forms or declinations of a verb as the same term (Harrower and Brewer, 2003). For example, stemming will ensure that the model, modelling, and modelled will be recognised as the term model. In this way, stemming will reduce the number of distinct terms and increase the frequency of some terms (Polon, 2011). Stemming has two common types (Mahgoub et al., 2008; Polon, 2011):

- ✓ Inflectional stemming: This aims to regularize grammatical variants such as present/past and singular/plural (this is called morphological analysis in computational linguistics). The degree of difficulty varies significantly from language to language.
- ✓ Stemming to the root: This aims to reach a root form with no inflectional or derivational prefixes and suffixes, which may lead to the least number of terms.

#### Task 3

In task 3, a numeric two-dimensional matrix representation of the corpus is created. Generation of the first form of the TDM includes three steps:



1. Specifying all the documents as rows in the matrix.
  2. Identifying all the unique terms in the corpus (as its columns), excluding the ones in the stopword list.
  3. Calculating the occurrence count of each term for each document (as its cell values)
- If the corpus includes a rather large number of documents as is commonly the case, then it is common for the TDM to have a very large number of terms (Li et al., 2010; Hill and Lewicki, 2007). Processing such a large matrix might be time consuming, and, more importantly, it might lead to extraction of inaccurate patterns. These dangers of large matrices and time-consuming operations pose the following two questions:
- What is the best representation of the indices for optimal processing by text mining programs?
  - How can the dimensionality of this matrix be reduced to a more manageable size to facilitate more efficient and effective processing?

To answer these questions, the various forms of representation of indices should be evaluated. One approach is to transform the term frequencies. Once the input documents are indexed and the initial term frequencies by document have been computed, a number of additional transformations can be performed to summarize and aggregate the extracted information (Diligenti et al., 2000). Raw term frequencies reflect the relative prominence of a term in each document. Specifically, terms that occur with greater frequency in a document could be the best descriptors of the contents of that document. However, it is not reasonable to assume that the term counts themselves are proportional to their importance as descriptors of the documents. For example, even though a term occurs three times more often in document A than in document B, it is not necessarily reasonable to conclude that this term is three times as important a descriptor of document B as it is for document A (Diligenti et al., 2000; Francis, 2003; Hill and Lewicki, 2007). In order to have a more consistent TDM for further analysis, these raw indices should be normalized. In statistical analysis, normalization consists of dividing multiple sets of data by a common value in order to eliminate different effects of different scales among data elements to be compared. Some of the normalization methods include: log frequencies; binary frequencies; and inverse document frequencies. The latter of these is the most commonly used transformation method (Mahgoub et al., 2008; Diligenti et al., 2000).

Another important aspect is to reduce the dimensionality of the TDM because it is often very large and rather sparse, with most of the cells filled with zeros. Some of the options available for reducing such matrices to a manageable size include (Hill and Lewicki, 2007):

- ✓ A domain expert goes through the list of terms and eliminates those that do not make much sense for the context of the study.
- ✓ Eliminate terms with very few occurrences in very few documents.
- ✓ Transform the matrix using singular value decomposition (SVD). SVD is a method of representing a matrix as a series of linear approximations that expose the underlying meaning-structure of the matrix. The goal of SVD is to find the optimal set of factors that best predict the outcome. During data preprocessing prior to text mining operations, SVD is used in latent semantic analysis (LSA) to find the underlying meaning of terms in various documents.

#### **4.4.3 Activity 3: Extract the knowledge**

Novel patterns are extracted in the context of the specific problem being addressed, using the well-structured TDM, and possibly augmented with other structured data elements (such as numerical and/ or nominal variables, potentially including the time and place specifications of the documents). These are the main categories of knowledge extraction methods in text mining studies (Diligenti et al., 2000; Li et al., 2010):

- ✓ Prediction such as classification, regression, and time-series analysis
- ✓ Clustering including segmentation and outlier analysis
- ✓ Association including affinity analysis, link analysis, and sequence analysis
- ✓ Trend analysis

## CHAPTER 5: FINDINGS AND DISCUSSION

The findings of the assessment of the perceptions on AI; citizen science study; confirmatory survey; and big data analytics (combining the different aspects) are presented and discussed in this chapter.

### 5.1 PERCEPTIONS OF WATER EXPERTS

It is no doubt that AI is infiltrating many disciplines and its impacts and that of big data analytics in general cannot be underestimated. One of these areas is the water sector. Thus, some water experts (19) were approached for their opinions and understanding of AI as it relates to this field. They were also interviewed regarding their views on the collection and use of citizen science data alongside the conventional water analytical data. The water experts included: academics and researchers (5 respondents); water utility companies (4 respondents); DWS (3 respondents); municipality officials working in the water sector (5 respondents); and environmental consultants (2 respondents).

Their responses are presented in Tables 5.1 and 5.2.

**Table 5.1. Responses of water experts on big data analytics and artificial intelligence (AI)**

Questions	Responses			
	Yes	No	Not sure	Notes
Do you know what big data and artificial intelligence are?	17		2	The 2 respondents (1 male and 1 female) that were not sure work in a municipality. They are matric holders and both over 50 years old.
Is big data analytics useful for water management?	17		2	The 2 are the same respondents mentioned above.
Do you have any reservations about AI?	6	13		Those apprehensive include 2 academics (both male and over 50 years old) and 4 municipal officials (incl. the 2 mentioned above). The reservations of the academics were generally around a fear of over empowering robots, thus losing the human aspect in tasks. The municipal officials indicated a fear of job losses if robots were to take over.
<p><i>Profiles of respondents:</i></p> <p>Municipalities: 4 females (1 with BSc Hons; 2 with BSc degrees and all in the 35-50 age group; 1 with matric above 50 years) and 1 male (with matric and above 50 years).</p> <p>Water utility companies: 1 female (MSc and aged 25-35 years) and 3 males (all with MSc and in the age group 35-50).</p> <p>Academics: 2 females (PhD and aged 35-40 years) and 3 males (PhD and aged 45-55 years).</p> <p>DWS officials: 1 female (MSc and aged 35-40 years) and 2 males (MSc and aged 35-40 years).</p> <p>Environmental consultants: 2 females (1 MSc, 1 PhD and aged 30-35 years)</p>				

There was a general understanding of big data analytics and AI and their implications in the modern-day workspace. The potential use of these techniques in water management was appreciated although there were fears and apprehension around the likely over dependence on AI that could lead to stifling the human aspect and natural intelligence. The other fear was of potential job losses and examples were made of the banking, call centres and retail sectors that are already shedding jobs in favour of robotic systems. It was clear that most jobs that will likely be lost are those with repetitive tasks that could be taught to robots. The consensus was that the human aspect and subject matter expertise (e.g. expertise in water) were not replaceable.

**Table 5.2. Responses of the water experts on their perceptions of citizen science**

Questions	Responses			
	Yes	No	Not sure	Notes
Do you know what citizen science is?	19			All respondents showed knowledge of what citizen science is. Others even gave common examples of citizen science such as people taking videos with their cellphones of incidents and sending them to news outlets for reporting.
Have you used citizen science directly/indirectly in your job?	5	14		4 municipal officials and one environmental consultant indicated that they had used or made use of or been exposed to citizen science. This was in cases where they got information (mostly complaints) from communities about the poor quality of the water supplied or its unavailability.
Do you think citizen science is beneficial?	16		3	The respondents who were not sure were 2 academics and 1 from a water utility company. Their emphasis was on rather using more scientific methods to define water quality, e.g. analytical techniques. This was in contrast with those who cited that citizen science could complement analytical techniques where the latter are not available. There was also some doubt in the veracity of citizen science in some cases.
Are there any barriers to creating this benefit?	9		10	Those who cited barriers included all the municipal officials; the environmental consultants; and DWS officials who indicated that political interference could be a problem. In their view, community engagement is usually viewed negatively by political authorities as there is fear that it can cause public panic. They also indicated that the cause is perhaps that most such engagements are done by activists, leading to protests if there are problems.
Do you think these barriers can be overcome?	9		10	The same 9 officials cited above indicated a number of ways that the barriers could be dealt with, including: engaging with communities when things are fine, i.e. not waiting for problems to occur or when it is

Questions	Responses			
	Yes	No	Not sure	Notes
				time for political campaigns. They said that this way, it would be easy to gauge the mood of the community if things were to go out of hand. So, a constant follow up on citizen science will help in building trust and the political authorities will have less to worry about.
Do you think citizen science can increase public scientific literacy?	19			All respondents indicated that community engagement is likely to increase their scientific literacy. The DWS officials and academics gave examples of communicating climate change through community engaging and how this would help them in understanding the changes that they have observed over time in their surroundings.
Do you think that citizen science can play a role in environmental stewardship?	19			All respondents indicated that community engagement was likely to increase environmental stewardship. The municipal officials gave an example of encouraging communities to sort out waste at their homes and how that could reduce the waste footprint and costs of sorting, making the money available for other important service delivery aspects. Another example was that of preventing erosion by showing the community how to create terraces and plant trees, etc.
Do you think that citizen science can instill environmental democracy?	19			There was consensus about the potential that citizen science has in instilling environmental democracy. The common view was that the communities had rights to take authorities to task regarding their environment and also to validate information released by the authorities. One example that was given was that of the pollution of the Vaal River due to sewage input. Other examples included pollution by companies operating within the communities that seemed to get off lightly with serious environmental violations.
Do you foresee citizen science being widely accepted by big data scientists?	16	1	2	The reasons around the veracity of citizen science were raised by 2 academics and 1 water expert from a water utility company. The majority of the other respondents thought that there was potential to tap into citizen science.
Would you be willing to incorporate citizen science data into conventional	13		6	Some respondents who had supported the acceptance of citizen science data in big data analytics seemed not sure if they would

Questions	Responses			
	Yes	No	Not sure	Notes
water quality data in your role as a water professional?				actually use citizen science in their own work. These included 3 academics; 1 municipal official; and 2 water utility officials. The skepticism was still around the veracity and possible bias in such data. One academic indicated that some communities may downplay some possible deleterious effects of water not deliberately, but perhaps as result of them getting used to the effects and developing some form of resistance. This last point is actually what our aim of conducting our own confirmatory surveys in the study was. It was to identify any biases that could be in the communities' citizen science data. The other respondents felt that incorporating citizen science into their conventional data would likely add value and this could make up for missing data.

Generally, there was consensus that citizen science data is important and that it would be beneficial to incorporate it into conventional data. However, the concerns raised were mainly around the veracity of such data and biases that may be inherent in it. Political interference was cited as a threat to the democracy of such data and was likely to lead to bias in some cases.

## 5.2 CITIZEN SCIENCE

A general survey was conducted to assess the opinions of the respondents of certain aspects. This is presented in Table 5.3.

**Table 5.3. General assessment of community respondents' opinions**


1. On a scale of <b>1 (Strongly agree)</b> to <b>6 (Strongly disagree)</b> , please rate your agreement with the following statements						
	<b>1</b>	<b>2</b>	<b>3</b>	<b>4</b>	<b>5</b>	<b>6</b>
a) Most people are basically honest	26	15	16	3		
b) Most people are trustworthy	17	18	21	4		
c) Most people are basically good and kind		36	24			
d) Most people are trustful of others		27	33			
e) I am trustful	60					
f) Most people will respond in kind when they are trusted by others	27	33				
2. On a scale of <b>1 (trust their information very much)</b> to <b>6 (do not trust their information at all)</b> , please tell me to what extent you trust the following sources of information to tell you the truth about the quality of water of your area.						
	<b>1</b>	<b>2</b>	<b>3</b>	<b>4</b>	<b>5</b>	<b>6</b>





a) Workers at the supplying authority's water treatment facility		15	33	6	6	
b) Scientists working for government agencies		15	33	12		
c) University scientists doing research in water		38	22			
3. On a scale of <b>1 (Strongly agree)</b> to <b>6 (Strongly disagree)</b> , please rate your agreement with the following statements						
	<b>1</b>	<b>2</b>	<b>3</b>	<b>4</b>	<b>5</b>	<b>6</b>
a) All things considered, the world is better off because of science and technology.		31	20	9		
b) Scientists know best what is good for the public.		36	17	7		
c) Water quality data collected by citizens adds value to data collected by scientists.	56	4				
d) Water quality data collected by citizens can be useful for decision-making where scientific data is not available.	55	5				
e) Scientists should incorporate information collected by citizens, e.g. colour and taste in determining water quality.	56	4				
f) I feel informed about the quality of the water I use for daily purposes.			47	6	7	
<b>Profiles of respondents:</b> Female: 30 Education level: 27 (completed matric); 3 (have no matric) Employment: 5 (teachers); 3 (nurses); 2 (security guards); 1 (tavern owner); 2 (work in lodges); 10 (do temporary jobs, receive children's grants); 7 (dependent on grants) Age groups: 4 (18-25 years); 10 (26-35 years); 9 (35-45 years); 4 (45-55 years); 3 (over 55 years) Male: 30 Education level: 21 (completed matric); 9 (have no matric) Employment: 3 (teachers); 3 (security guards); 2 (tuck shop owners); 2 (employed at a filling station); 2 (work in lodges); 16 (do temporary jobs); 2 (pensioners) Age groups: 7 (18-25 years); 12 (26-35 years); 6 (35-45 years); 3 (45-55 years); 2 (over 55 years)						

**Table 5.4. Assessment of the community respondents' knowledge about their water**

Questions	Responses			
	Yes	No	Not sure	Notes
Do you get water from a borehole as your main source of water?	51		9	Those not sure were mainly urban dwellers who were not sure where the municipality water they receive actually comes from. Four of them indicated that they used bottled water for drinking.
Do you use a community borehole?	45	6	9	Six of those adamant that they are not using a community borehole had boreholes on their properties. The other 9 are the same from above that were not sure about the source of their water. Those using community

Questions	Responses			
	Yes	No	Not sure	Notes
				boreholes indicated that they used water containers to fetch this water from community borehole taps, boreholes in schools, clinics or community holes. These sources were usually near their homes.
In the past two weeks, was the water from this source not available for at least one full day?	6	54		Those that indicated that water was not available for a full day said that it was due to a malfunctioning borehole pump.
Do you consider the water you drink from the borehole or tap to be safe for consumption?	48	4	8	The 4 that were adamant that the water was unsafe are the same ones that said they use bottled water. The other 8 people, most in Mahikeng were unsure about the safety of their water although they consumed it anyway.
Does the water taste fine for you?	56	2	2	Those using bottled water were either unsure of the taste (because they did not use it for drinking) or indicated that the last time they tasted it was salty and hard.
Does the water have a bad smell?		60		Everyone indicated that the water had no foul smell.
Is the water colour normal, clear?	54	6		<p>Six people in Mahikeng said that at times the piped municipality water is cloudy. One respondent showed us the picture below that they took at one instance when the water was cloudy. They did not use this water until it became clear. From our observation, this could be due to the chlorination process.</p> 
Do you treat the water in any way to make it safe for drinking?	21	39		Those who indicated that they treat the water for drinking

Questions	Responses			
	Yes	No	Not sure	Notes
				<p>purposes indicated that they boil the water. A few said they add some drops of bleach (Jik). However, the majority (including the 4 that only use bottled water) said that they do not treat it. One of the respondents from Mahikeng shared a picture of some filters that a relative of theirs (not interviewed) uses to treat their water.</p>  <p>The 3 packs show the stages of use with the most used filter on the left and the least used on the right. The brown colouration is likely to be iron. It is not conclusive whether this could be cumulative from the low concentrations (as determined in our confirmatory survey) in the water or it could have been from corrosion of metal components.</p>
Does the water cause formation of a dry skin after washing with it?	60			<p>Everyone indicated that their skin became dry after bathing with the water. One respondent demonstrated this for us in Mahikeng (see photo below). The water was said to form a poor lather with common laundry bars and soap. Our own confirmatory survey showed that most of the water causes dry skin and that the effect was more prominent for the water in Mahikeng. Some indicated that they use a dishwasher to make the water smoother. This has the same effect as a shower gel.</p>

Questions	Responses			
	Yes	No	Not sure	Notes
				
Would you be willing to share your assessments of water quality and safety with water treatment and supplying authorities?	60			All the respondents showed a willingness to share issues regarding their water with the authorities.

### 5.3 CONFIRMATORY SURVEY

The results for anion and cation levels in the water collected during the confirmatory survey are shown in Table 5.5. The WQI and total hardness have also been determined.

**Table 5.5. Anion and cation levels in groundwater (nd – not detected; ppt – parts per thousand)**

ID	Area	Latitude	Longitude	pH	EC	TDS	Cl <sup>-</sup>	NO <sub>3</sub> <sup>-</sup>	Ca	K	Mg	Na	P	Si	WQI	Total Hardness
					$\mu\text{S cm}^{-1}$	ppt										
							$\text{mg l}^{-1}$									
ZMQRS1	Zeerust	-25,54083	26,07972	7.28	610	300	6.720	2.211	73.57	0.36	36.34	2.297	0.1	6.29	Excellent	Hard water
Malebelele CPA	Malebelele Lekubu	-25,23325	26,0512	7.44	480	240	nd	11.83	72.54	0.09	35.77	nd	0.07	7.56	Excellent	Hard water
Primary School	Malebelele Lekubu	-25,3697	26,11008	7.75	340	170	11.04	39.23	94.38	0.09	21.79	nd	0.07	2.59	Excellent	Hard water
Mokgola Clinic	Mokgola	-25,1945	26,7231	7.2	530	260	7.254	0.221	274.82	0.14	16.22	nd	0.08	7.57	Good	Hard water
Garage	Montsana	-25,2767	26,92326	7.15	800	400	8.40	3.923	229.58	0.09	50.35	nd	0.09	8.42	Good	Hard water
Madikwe	Wonderboom Gate	-24,7672	26,1624	7.28	980	490	11.32	18.21	459.16	0.12	59.7	nd	0.09	1.85	Poor	Hard water
Nkosinathi liquor	Supingstad	-24,4647	26,61475	8.28	970	450	12.08	9.761	461.24	0.08	61.61	nd	0.08	3.57	Poor	Hard water
Supingstad tap	Supingstad	-24,4722	26,4167	7.57	900	450	12.37	9.741	471.9	0.09	60.79	nd	0.08	3.53	Poor	Hard water
Supingstad tap2	Supingstad	-24,47153	26,34208	7.35	740	370	12.76	12.05	56.42	0.09	65.43	nd	0.08	3.92	Excellent	Hard water
Supingstad Clinic	Supingstad	-24,7889	26,05338	7.18	980	490	12.28	10.33	100.62	0.13	5.78	nd	0.07	nd	Excellent	Hard water
SA-Botswana border	Kopfontein	-24,7079	26,1023	8.42	170	80	0.512	nd	363.74	0.09	60.11	nd	0.08	3.44	Good	Hard water
Villa Rosa Zeerust	Zeerust	-25,536	26,06625	7.56	540	270	7.176	1.913	320.06	0.09	27.23	nd	0.08	1.68	Good	Hard water
Private house	Magogoe-Mahikeng	-25,886557	25,600435	7.1	550	260	7.920	5.438	234.78	0.079	19.38	nd	0.08	1.37	Good	Hard water
GJ Podile School	Ramosadi-Mahikeng	-25,940337	25,6822287	7.02	1560	790	35.27	83.19	489.32	0.08	92.07	nd	0.07	5.76	Poor	Hard water
Ipeleng Primary School	Mahikeng-Montshiwa Unit 2	25,841223	25,615913	7.62	850	420	30.18	4.240	157.82	0.22	32.27	nd	0.08	0.26	Good	Hard water
Seetsele Primary School	Goo-ra-makgetla - Mahikeng	-25,841233	25,615913	8.0	77	390	28.74	6.257	193.44	0.21	32.61	nd	0.08	0.53	Excellent	Hard water
Motoseng	Motoseng-Mahikeng	-25,860779	25,586562	7.46	880	440	15.57	17.88	235.04	0.09	61.57	nd	0.08	5.84	Good	Hard water
Sedikwa Lodge	Mahikeng			7.66	430	210	5.122	6.39	258.7	0.07	20.5	nd	0.07	1.53	Good	Hard water
	WHO standard			6.5-8.5	500	500	200	50	75	12	50	200	10	17		

The following variables were analysed, but the results are not presented in the table: Fe,  $\text{SO}_4^{2-}$ ,  $\text{F}^-$  and Al. Their concentrations (in  $\text{mg l}^{-1}$ ) were in the ranges: Fe (0.01-0.03) and Al (0.01-0.02) while the others were not detected. Only 4 of the 17 samples were classified as poor according to the WQI. However, the water at these boreholes was found to be usable from our own assessment of taste, colour and smell and also from the citizen data provided by the communities. All the samples had a slightly salty taste, which is synonymous with most borehole water. They were classified as hard water as inferred from geochemical modelling. This was

confirmed by the elevated carbonate ( $\text{CaCO}_3$ ) concentrations determined in some of the samples (Table 5.6). Some of the photos in Fig. 3.1 show quite apparent effects of high alkalinity (and water hardness), e.g. in the scaling observed and precipitates on the banks of the Klein Maricopoort Dam. Some of these cumulative effects have been observed in toilet cisterns and some taps. Also, the high temperature of the water in storage tanks (in one instance it was around  $40^\circ\text{C}$ ) are known to increase scaling as carbonates tend to precipitate as temperature increases. That is why scaling easily occurs in boiling kettles, shower areas and other high temperature systems, e.g. pipes in cooling towers of thermal power stations. This precipitation trend is different for sulphates which tend to dissolve as temperature increases.

**Table 5.6. Confirmatory survey of alkalinity content in selected samples**

Sample	pH	Alkalinity (as $\text{mg l}^{-1} \text{CaCO}_3$ )	Classification criteria	Indication
Nkosinathi Liquor Store	8.28	218.75	>180	Very hard water
GJ Podile School	7.02	125	120 to 180	Hard water
Ipeleng Primary School	7.62	93.75	60 to 120	Moderately hard water
Motsoeneng	7.46	143.75	120 to 180	Hard water
Sedikwa Lodge	7.66	68.75	60 to 120	Moderately hard water

## 5.4 DATA ANALYTICS

Advanced data analytics was performed using the dataset provided by NMMM that had most of the data for other areas outside the aquifer zone. This was used for transfer learning to approximate the missing cases in the IGRAC database (the one covering the aquifer area).

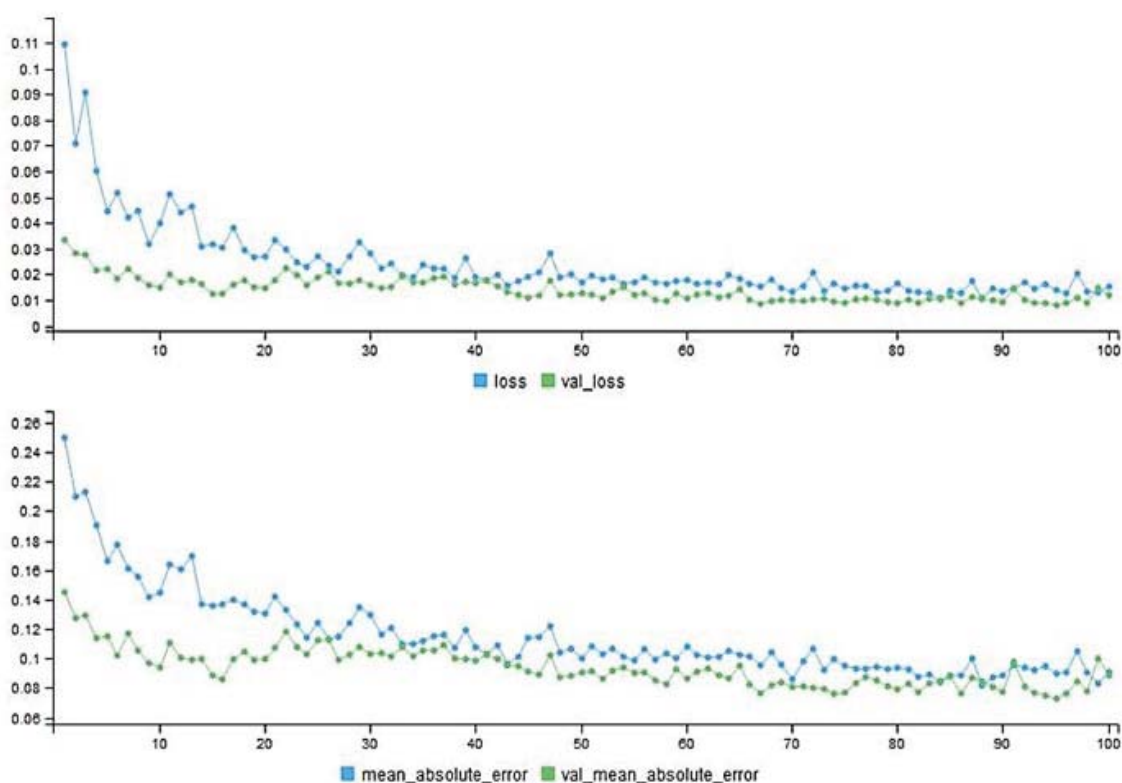
### 5.4.1 Imputation of missing values

The optimum multilayer perceptron topologies for Keras for efficient imputation of missing values for groundwater samples were 7:960:1, which expresses the: numbers of neurons in the input, hidden layers, and output layers, respectively. Table 5.7 summarises the performance of the model when approximated with a single variable of missing data. All the approximations show efficiency in the imputation of the variables such as  $\text{Ca}^{2+}$ ,  $\text{K}^+$ ,  $\text{Mg}^{2+}$ ,  $\text{Na}^+$ ,  $\text{SO}_4^{2-}$ ,  $\text{Cl}^-$ ,  $\text{CaCO}_3$  and pH. This is indicated by the high correlation coefficients; a low root means square error (RMSE) and high accuracy (%) associated with the variables.

**Table 5.7. Accuracy results for different variables from modelling with Keras**

Variables	No. of hidden nodes	Accuracy(%)	r	RMSE
Ca <sup>2+</sup>	100	81.72	80.92	0.006
K <sup>+</sup>	150	79.61	87.67	0.001
Mg <sup>2+</sup>	120	80.9	82.45	0.003
Na <sup>+</sup>	100	83.45	89.56	0.0007
SO <sub>4</sub> <sup>2-</sup>	100	82.16	85.67	0.002
Cl <sup>-</sup>	120	84.53	86.72	0.001
CaCO <sub>3</sub>	130	75.62	84.23	0.004
pH	120	81.42	85.67	0.003

After assessing the performance of the current model, 109 cases with missing values were tested to approximate the missing values for Ca<sup>2+</sup>, K<sup>+</sup>, Mg<sup>2+</sup>, Na<sup>+</sup>, SO<sub>4</sub><sup>2-</sup>, Cl<sup>-</sup>, CaCO<sub>3</sub> and pH. The estimation was assessed by imputing each of the eight attributes individually. In all, 1089 training steps were used. The new dataset of 497 cases was restructured and used for further assessment. The error minimisation during the training is shown in Fig. 5.1.

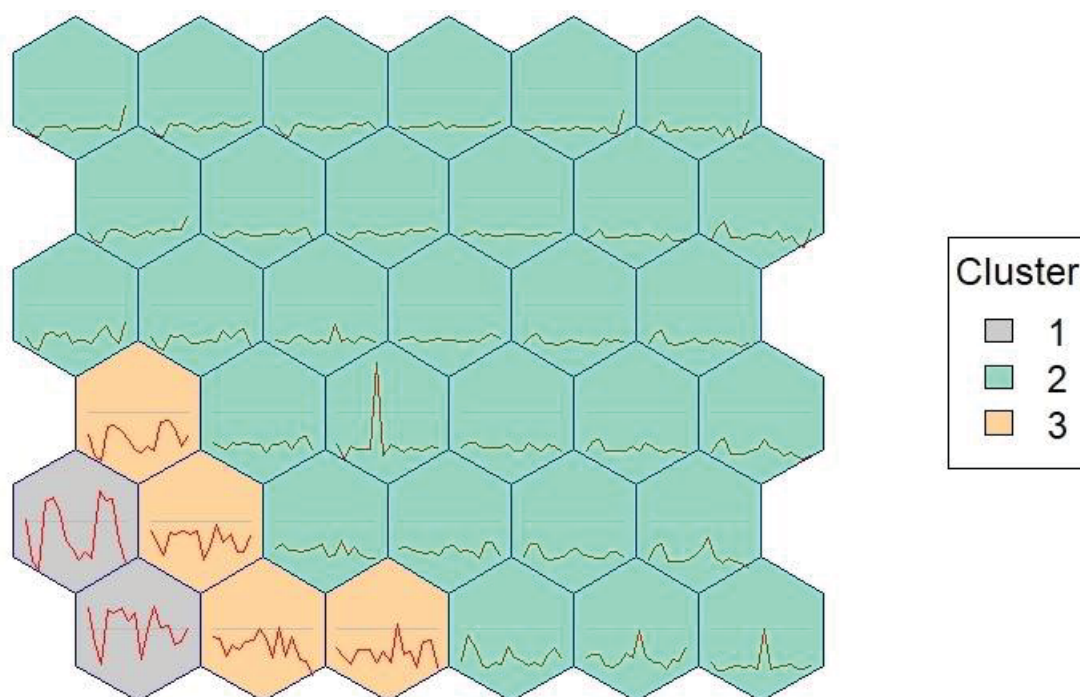
**Figure 5.1. Error minimisation during training of neural network**

#### 5.4.2 Self-organising maps for hydrochemical assessment

After transfer learning, the two complete datasets were then used for further analytics, e.g. clustering. The k-means-SOM hybridisation allowed for the grouping of groundwater samples into three clusters as presented in Fig. 5.2. In general, the order of the enrichment in most of the hydrochemical variables followed the order:



cluster 2 > cluster 3 > cluster 1. Fifty five percent, 30%, and 15% of the samples were grouped in cluster 2, cluster 3 and cluster 1, respectively. Cluster 2 and cluster 3 consisted of borehole samples with the concentration of the hydrochemical variables lower than the threshold permitted limit set by WHO (2004). These are classified as excellent and good water types and are appropriate to be used as drinking water. The last cluster (cluster 1) contains borehole samples with a concentration of hydrochemical variables higher than the threshold permitted limit, making the water in that cluster to be poor and unsuitable for drinking.

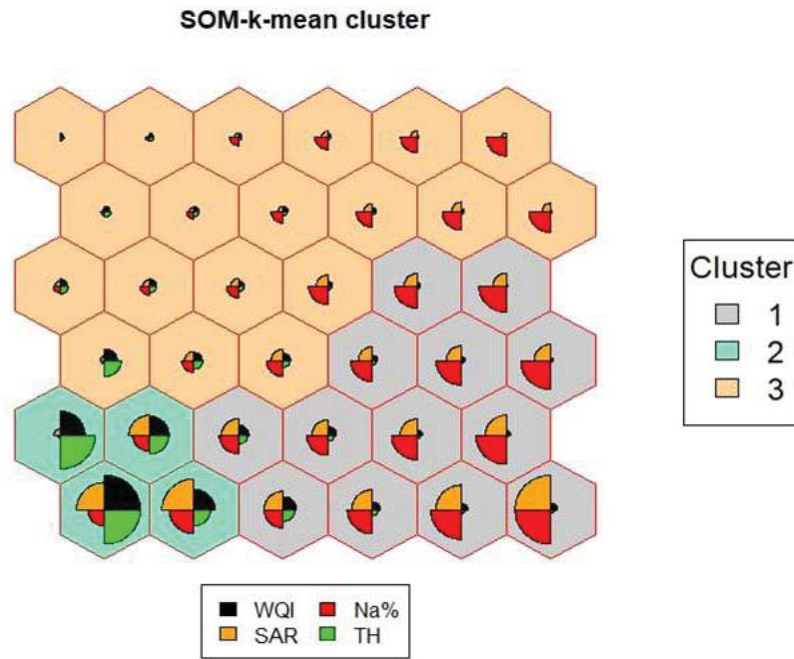


**Figure 5.2. Clustering of borehole water using a self-organising map**

It should be noted that within each cell are some samples and the variable quantities in them are shown by the squiggle lines. These numbers vary depending on the proximity of the samples. Within each cell, there is a “winner” cell around which neighbours with similar properties arrange. The cells then aggregate to form a similar cluster which is separated from other different clusters by distinguishable borders.

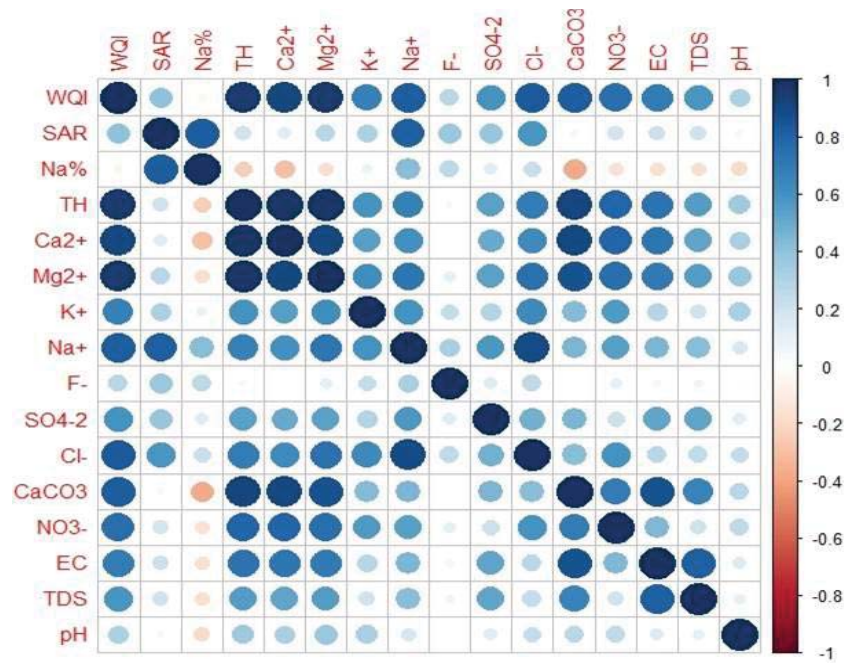
#### 5.4.2.1 Assessment of groundwater quality for drinking purposes

The full chemical analyses of groundwater that were used for quality assessment and comparison with the WHO (2004) guidelines for drinking water are summarised and available in the link ([https://www.dropbox.com/sh/jo7meb7fghnp01c/AABAxqlbpJ02SGm5A\\_chdRdea?dl=0](https://www.dropbox.com/sh/jo7meb7fghnp01c/AABAxqlbpJ02SGm5A_chdRdea?dl=0)). The computed WQI values range from 0-50, 50-100, 100-300 for cluster 3, cluster 1 and cluster 2, respectively. The results of WQI indicate that 55% of the groundwater samples in grouping cluster 3 were in the class of excellent water quality, while 30% and 15% in first and second clusters were in the category of good and very poor water quality, respectively (Fig. 5.3). The quality of the groundwater in cluster 2 can be due to effective leaching of ions, overexploitation of groundwater, direct discharge of effluents, or unimproved sanitation.



**Figure 5.3. k-means-SOM clustering for WQI, SAR, %Na and total hardness (TH).**

Variables such as  $\text{Ca}^{2+}$ ,  $\text{Mg}^{2+}$ ,  $\text{K}^+$ ,  $\text{Na}^+$ ,  $\text{SO}_4^{2-}$ ,  $\text{Cl}^-$ ,  $\text{CaCO}_3$ ,  $\text{NO}_3^-$ , EC and TDS show a strong correlation with WQI (Fig. 5.4). The level of relationship that exists between variables is calculated by Pearson's correlation coefficient. The higher the value of the correlation coefficient, the better and more beneficial the association (Sanchez-Martos et al., 2002). Also, a high positive correlation was observed between:  $\text{Ca}^{2+}$  and  $\text{Mg}^{2+}$ ;  $\text{Ca}^{2+}$  and  $\text{CaCO}_3$ ;  $\text{Ca}^{2+}$  and  $\text{NO}_3^-$ ;  $\text{NO}_3^-$  and  $\text{Mg}^{2+}$ ;  $\text{Mg}^{2+}$  and  $\text{CaCO}_3$ ;  $\text{Na}^+$  and  $\text{Cl}^-$ ; EC and  $\text{CaCO}_3$ ; TDS and EC. The high positive correlation between the variables could be an indication that cations or anions might have the same origin. These correlations imply that the dominant elements, which contribute to the salinity of groundwater and the tendencies among them, follow a similar trend. For Ca and Mg, it is quite plausible as the main constituent aquifer rock is dolomite.



**Figure 5.4. Correlation matrix for the variable interrelationships**

#### 5.4.2.2 Ion ratio coefficients

Based on the exploration of the dataset, Ca and Mg are the significant cations and  $\text{Cl}^-$  is an important anion in the study area. The presence of the dolomite rock ( $\text{CaMg}(\text{CO}_3)_2$ ) is the likely source of Ca and Mg and hardness in the groundwater. The ratio of Ca/Mg shows that around 65% of the boreholes have a ratio Ca/Mg ranged from 0 to 1, which indicated that the contribution of dolomite dissolution on Ca and Mg levels in the water samples. About 27% of the samples have a Ca/Mg ratio higher than 2, which also indicated the dissolution of silicate minerals in addition to the Ca/Mg concentration in groundwater samples. The samples showing a Ca/Mg ratio between 1 and 2, which shows carbonate rock dissolution on Ca and Mg concentration in groundwater samples showed a high TH (Fig. 5.3). These results showed that 85% of the borehole samples grouping in cluster 3 and 1 are classified as hard and very hard water whereas only 15% in the second cluster are sorting as very hard water.

Moreover, salinity in the groundwater can come from different sources, which could be identified by the correlation between the  $\text{Na}^+$  and  $\text{Cl}^-$  in the groundwater. Hence, the average molar ratio of Na/Cl is 1.83 in the study area, which indicates higher  $\text{Na}^+$  values than  $\text{Cl}^-$ . This indicates the role of evapotranspiration and evaporation in the increased sodium levels in the groundwater samples.

#### 5.4.2.3 Suitability of the water for irrigation

While of more importance is the potability of the water for the communities, its suitability for agricultural use is also important. Some gardening activities were observed during the visits and these made use of the water from the boreholes. Groundwater quality was assessed for its suitability for irrigation by determining parameters such as sodium percent (%Na), SAR and EC (Belkhiri and Mouni, 2012). Electrical conductivity (EC) and TDS are a good measure of salinity in water. The salinity hazard is directly related to the concentration of ions in the water, which also affects soils by enriching with a sodium concentration when

sodium values are much higher than calcium in the water. It can destroy the structure of soil due to the dispersion of clay particles and decrease the osmotic activity of plants. Usually, water with EC values greater than  $3000 \mu\text{S cm}^{-1}$  is considered to be of unsuitable quality for irrigation purposes. All the boreholes showed EC values below  $1000 \mu\text{S cm}^{-1}$ , which is classified as good (55%) and permissible (45%) for agricultural usage. The concentration of sodium, calcium, and magnesium, and sodium in water can influence the normal infiltration rate of the water. The alkali/sodium hazard to crops is measured by the SAR. In this investigation, the SAR results (Fig. 5.3) for the samples in cluster 3, cluster 1 and cluster 2 were found to be excellent, good and moderate for agriculture, respectively. The results for %Na showed that 55% (cluster 3) of the samples had been classified as good while 45% (cluster 2 and 1) of the samples were classified as permissible for agriculture, respectively.

### 5.4.3 Text mining for citizen science

The quantitative dataset for water quality of boreholes in NMMM (Mafikeng) and RTA (whose data was partly from IGRAC as well as prediction using transfer learning) were pre-processed in order to convert their numerical data to text. To conduct data pre-processing, the alkalinity of each borehole was calculated. The obtained alkalinity values were used to categorise the borehole (according to the quality of water for drinking and usage by households). The comments from people regarding their perceptions about water quality during the interviews (i.e. citizen science) were used to associate to match with each range of the alkalinity. The following ranges were defined:

- Alkalinity from  $0\text{-}59 \text{ mg l}^{-1}$  was defined as water of excellent quality, not salty and which did not need the use of too much soap to wash clothes and dishes.
- Alkalinity from  $60\text{-}120 \text{ mg l}^{-1}$  was defined as normal water, while sometimes it can taste slightly salty.
- Alkalinity above  $120 \text{ mg l}^{-1}$  was categorised as salty water, needing a lot soap or detergent to wash clothes and dishes. It also left scaly residue after washing hands, clothes and glasses.

The coding for text mining was conducted in R. An example of how this was conducted is presented (Table 5.8). In the example, an R code was used to convert the alkalinity range values to people's perception on the quality of water (comments). These comments were coded with respect to the alkalinity of each borehole. After converting the numerical data to text data, the comments were then rated as positive, neutral or negative. The positive, neutral and negative rates were associated with the alkalinity ranged of  $0\text{-}59$ ,  $60\text{-}120$  and greater than  $120$ , respectively.

**Table 5.8. Text data obtained after converting numerical data**

Boreholes	Alkalinity (mg l <sup>-1</sup> )	Comments from the alkalinity values	Rate
10-77258	0.02	The water is good, not salty, I don't use too much soap to wash clothes and dishes	Positive
10-78105	2	The water is good, not salty, I don't use too much soap to wash clothes and dishes	Positive
10-77266	7.01	The water is good, not salty, I don't use too much soap to wash clothes and dishes	Positive
23-00002	6.5	The water is good, not salty, I don't use too much soap to wash clothes and dishes	Positive
10-77267	6.8	The water is good, not salty, I don't use too much soap to wash clothes and dishes	Positive
10-77265	8.6	The water is good, not salty, I don't use too much soap to wash clothes and dishes	Positive
20-00039	10.2	The water is good, not salty, I don't use too much soap to wash clothes and dishes	Positive
10-77336	11.2	The water is good, not salty, I don't use too much soap to wash clothes and dishes	Positive

The approach used was to come up with a text dataset which can be applied for any classification modelling. The text data obtained could then be used for mining and used predictive modelling for text classification and prediction outcomes. The probability distributions of negative, neutral and positive rates were 10%, 32% and 58%, respectively. The complete dataframe and explanatory notes for the text mining are available in the Dropbox link: <https://www.dropbox.com/sh/oqa6xdtdf916h9/AABEiKUCVvJzRju7ae2O291Ba?dl=0>

#### 5.4.3.1 Converting text to digital format

There are many packages in the R ecosystem for performing text analytics. One of the newer packages is *quanteda*. The *quanteda* package has many useful functions for quickly and easily working with text data. It applies a function called *tokens* as the iterator winds its way through the entire object. This means that an iterator can apply functions to large objects that may be too large to fit in memory. Here *tokens* iterate through *train.text* and wraps *str\_split* to separate individual words as shown in the examples below.

```
# train.tokens[[357]]
```

```
[1] "water" "is" "good" "nonsalty" "I" "dont" "use" "to" "much" "soap" "to" "wash"
```

```
[13] "clothes" "and" "dishes"
```

Along the way, *tokens\_select* is applied along with another function *tokens\_wordstem*. The *tokens\_wordstem* function perform stemming on the tokens by removing the stopwords. The transformation would be as follows:

```
# train.tokens[[357]]
```

```
[1] "water" "good" "nonsalty" "dont" "use" "much" "soap" "wash" "clothes" "dishes"
```

The words such as “I” and “to” are considered stopwords and are removed from the previous *train.tokens* data set.



*First bag-of-words model*

Bag-of-words treats every word or groups of words, called n-grams as a unique feature of the document. Word order and grammatical word type are not captured in a bag-of-words analysis. As result, bag-of-words create a matrix dataset which fits into machine learning frameworks because it provides an organized matrix of observations and attributes. As result, the train.tokens.dfm data frame obtained has in each row document or individual corpus. The train.tokens.dfm columns are made of words or word groups. The word or word groups are the rows while the documents are the columns (Table 5.9). Train.tokens.dfm was converted to train.tokens.matrix to fit well in the machine learning model (the codes are available in the Dropbox link given above).

**Table 5.9. Bag-of-words model**

Document	water	good	nonsalty	dont	use	much	soap	wash	clothes	dishes	normal
Text1	1	1	1	1	1	1	1	1	1	1	0
Text2	1	1	1	1	1	1	1	1	1	1	0
Text3	1	1	1	1	1	1	1	1	1	1	0
Text4	1	1	1	1	1	1	1	1	1	1	0

For instance, text1 was converted to 1 1 1 1 1 1 1 1 1 1 1 digits which can be read as “water is good, non-salty, I don’t use too much soap to wash clothes and dishes”.

*5.4.3.2 Cross validation*

For best model fitting, it is important to perform a cross validation (CV) as the basis of our modelling process. Using CV, we can create estimates of how well our model will do in production on new, unseen data (Table 5.10). CV is powerful, but the downside is that it requires more processing and therefore more time. In order to perform the CV, the rate column was added in the train.tokens.dfm data frame and saved as train.tokens.df (see the entire data frame and code in the Dropbox link). In the CV model process, caret was used to create stratified folds for 10-fold cross validation repeated as shown in the code below:

```
# cv.folds <- createMultiFolds(train$Rate, k = 10, times = 3)
```

```
# cv.cntrl <- trainControl(method = "repeatedcv", number = 10, repeats = 3, index = cv.folds)
```

Since the data frame is non-trivial in size. As such, CV runs take quite a long time to run. To cut down on total execution time, the doSNOW package was used to allow for multi-core training in parallel. A single decision tree algorithm was used as the first model.

```
# rpart.cv.1 <- train(Rate ~., data = train.tokens.df, method = "rpart", trControl = cv.cntrl, tuneLength = 7).
```

The results from this model were as follows:

872 samples

40 predictors

3 classes: 'Negative', 'Neutral', 'Positive'



No pre-processing

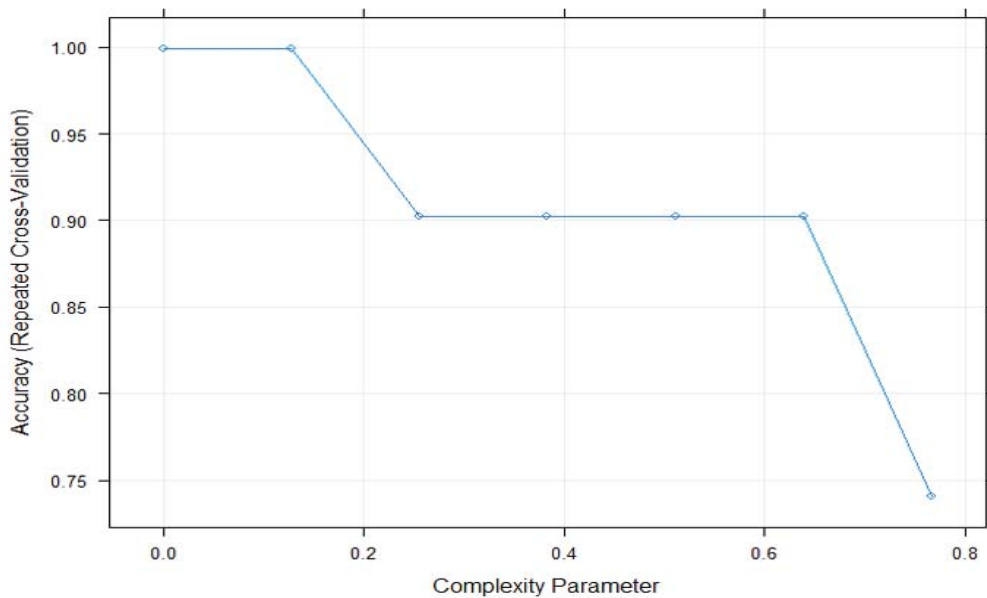
Resampling: Cross-Validated (10 fold, repeated 3 times)

Summary of sample sizes: 784, 785, 786, 785, 784, 785, ...

**Table 5.10. Resampling results across tuning parameters**

<b>Cp</b>	<b>Accuracy</b>	<b>Kappa</b>
0.000	0.9988505	0.9979066
0.1278539	0.9988505	0.9979066
0.2557078	0.9025517	0.8153267
0.3835616	0.9025517	0.8153267
0.5114155	0.9025517	0.8153267
0.6392694	0.9025517	0.8153267
0.7671233	0.7410084	0.4034050

Accuracy was used to select the optimal model using the largest value (Figure 5.5). A complexity parameter (Cp) of 0.13 with an accuracy and kappa of 0.99 and 0.99, respectively, were selected as the best set of the validation.



**Figure 5.5. Cross validation plot**

These results revealed that TextLength approach is very predictive and pushed the overall accuracy over the testing data to 100%. The power of cosine similarity can also be used to engineer a feature for calculating, on average, how alike each comment rate is to all of the comment rates. A confirmation that the positive comment rate has a high cosine similarity than negative and neutral comment rate is presented (Figure 5.6).

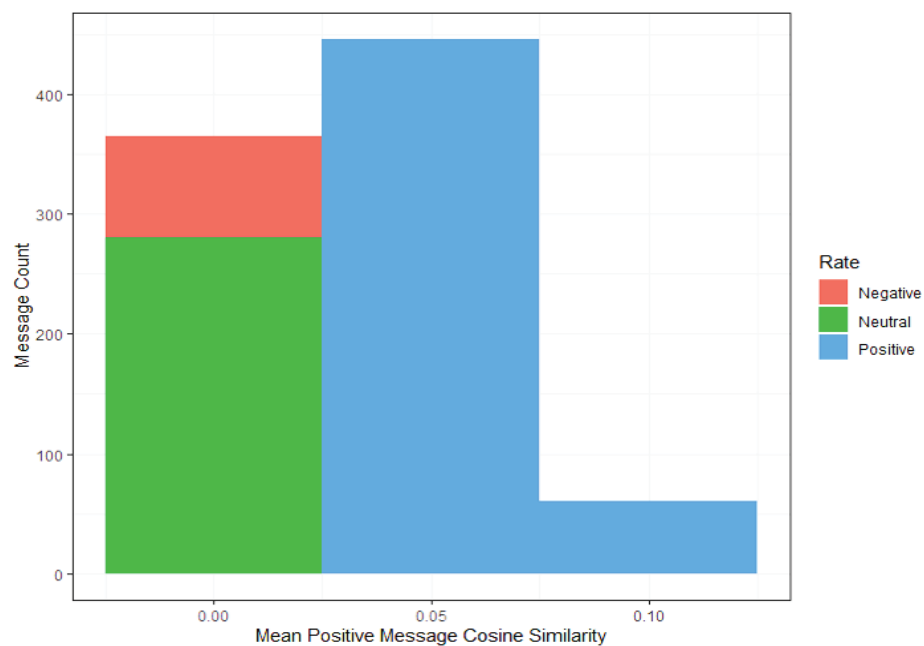


Figure 5.6. Distribution of comment rates using positive Cosine Similarity

## CHAPTER 6: REFLECTIONS ON LEARNING OPPORTUNITIES

---

*This chapter gives an overview of learning opportunities that the project opened for the research team and stakeholders. These include training programmes, workshops, community involvement and further collaborations.*

### 6.1 TRAINING PROGRAMMES

A MSc student (Lungisa Ngundu) was sent for a 3-month internship at IBM Africa Research during the project. The training included: machine learning, deep learning and text mining, making a significant contribution to the accomplishment of the project tasks related to these. This has also opened up opportunities for training of other students, not necessarily with respect to this project, but as a spinoff beyond it.

The research team also participated in webinars hosted by the US Geological Survey and the Sustainable Water Programme. These covered aspects such as: hydrogeology, citizen science, remote sensing and international law governing transboundary water resources.

### 6.2 WORKSHOPS

Some workshops were held internally for our research team and other students not in the research team. The workshops focused on training in machine learning and coding, skills that are still lacking in university curricula as they are viewed as belonging to computer science and engineering. Further, training on writing manuscripts was also offered.

Workshops for other stakeholders such as water utility companies have been discussed and will be offered at a later stage. There has been an increased demand of these, including by other universities as such skills have been viewed as only being limited to computer science and engineering.

In September 2019, our team conducted a one-day workshop on machine learning and coding in R for the Nuclear Research Centre in Belgium as part of a bilateral project that we have with them. This was followed by a seminar on “Machine learning for environmental applications” presented at the Nicolaus Copernicus University in Torun, Poland.

### 6.3 CURRICULUM DEVELOPMENT

One of the areas that the project intended to explore as part of its legacy was curriculum development. This involved developing a course on applications of AI, machine learning and coding specifically for chemistry. This had been prompted by the fact that no such a course existed that would bridge the gap for chemistry students wanting to pursue this field in combination with chemistry. A lot of changes in industrial production systems and research are taking place that have seen AI take a leading role, implying that graduates need to be better prepared.

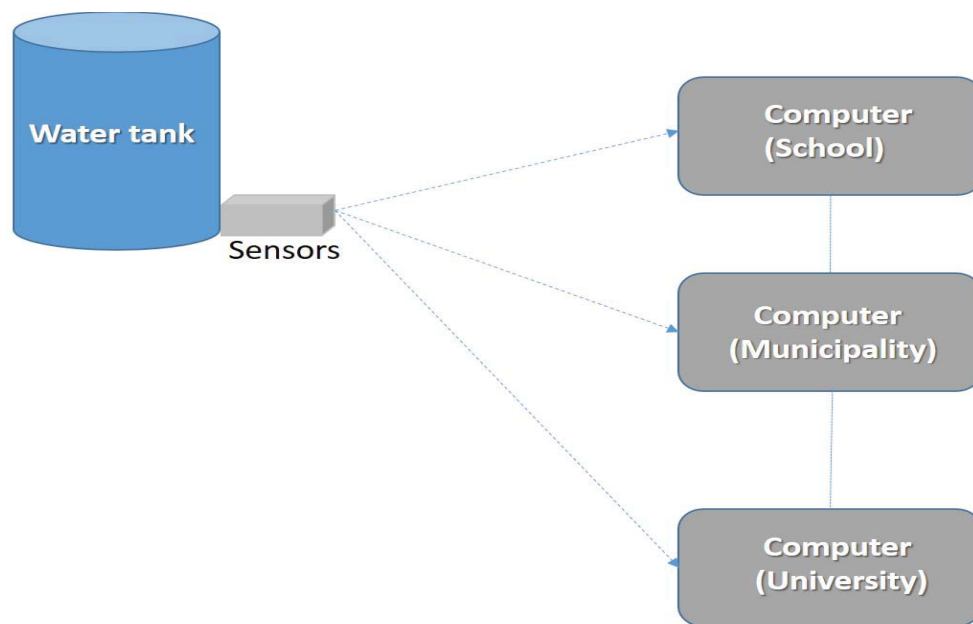
That course was developed and is being offered in 2020 as an elective module for the BSc Honours in Chemistry curriculum. It has drawn interest from several students as they perceive that it is part of the preparation that they need for the changing workplace. We are currently working on primers or tutorial books that can be used for the course in future and the big data tools reported in this study will feature prominently in them. This has also given a platform for the development of material for workshops to be conducted for stakeholders.

## **6.4 COMMUNITY INVOLVEMENT**

The communities that we interacted with expressed interest in participating in activities that ensure that they influence service delivery aspects in their areas. Their concerns were based on scepticism related to trust issues with politicians, fearing that their involvement in such endeavours, e.g. through citizen science may be misconstrued for political activity. Notwithstanding, there remains a lot of opportunity in using citizen science to tap into important information that would otherwise remain unused if conventional methods only are used. This has also been made easier to use by the availability of smart phones and user applications such as WhatsApp and platforms that allow for communication and capturing of information.

## **6.5 FURTHER COLLABORATIONS**

The project revealed a number of areas of collaboration with other interested parties and stakeholders to whom aspects of AI and big data analytics may be important. These areas include amongst others the internet of things (IoT), an increasingly important area that combines a variety of aspects such as sensor technology. Some of the organisations that have expressed interest in pursuing further applications of the findings of the project include SqwidNet (Johannesburg) and the Greater Giyani Municipality (Limpopo). The former supply Raspberry Pi gadgets (with virtual Wi-fi) that are important in sensor technology. This makes it possible to deploy sensors in the field, capture the data in real time, and distribute it to various computer stations (Figure 6.1).



**Figure 6.1. Layout of the water quality monitoring system involving internet of things (IoT)**

This example is a conceptualisation of the application of this technology to monitoring and assessing of drinking water quality for a school in certain municipality. With Raspberry Pi, it is possible to connect sensors (i.e. probes for pH, temperature, conductivity) to a virtual Wi-Fi where the measurements can be recorded in real time. The data can then be relayed to remote computers for use by laboratory technicians, water managers and researchers. The data can be collected at chosen intervals, e.g. hourly and relayed to remote computers that have algorithms established to run in batch mode (i.e. per tranche of data received) or in continuous mode (i.e. online processing of the data as it arrives). The purpose of the algorithms is to assess the patterns in the data received and to draw insights regarding the likely quality of the water in future and any changes that could occur. This way, educators and learners can be involved in aspects of AI that inform service delivery (water quality monitoring) within their community.

## CHAPTER 7: CHALLENGES ENCOUNTERED

---

While a number of accomplishments were made and learning opportunities observed, there were some challenges that were encountered in the project. These can be divided into: data-related aspects and stakeholder engagement.

### 7.1 STAKEHOLDER ENGAGEMENT

Challenges were encountered in dealing with some of the municipalities regarding general information on water issues and data acquisition specifically. For instance, where municipalities in urban areas had mixtures of groundwater and surface water there was no available information regarding ratios of such mixtures and whether this was conducted occasionally or was an established exercise. The research team could not access the water treatment sites to understand the physical and chemical processes used to prepare water for distribution and consumption.

There was also a challenge of establishing cooperation with government departments in relation to data-sharing. This took very long to be resolved, causing a loss of time in the process. Moreover, once the data had been accessed it was found to contain many missing values. This would have ideally been easier to deal with at the onset if the cooperation had been established earlier.

Engagement with community members was difficult in cases where they had to respond via their cellphones. This was largely due to lack of money to purchase data bundles and airtime. This resulted in some instances, e.g. video recordings of the water and sources not being taken. Ideally, it would have been helpful to have recordings of water properties such as colour changes as this would contribute to citizen science. While there exists opportunity to use platforms such as social media and smartphones to collect water data, the plight of unemployment and poverty in most communities is the prohibitive factor.

### 7.2 DATA-RELATED ASPECTS

As indicated above, a big challenge related to the data obtained was the prevalence of missing values. In most cases, this made the data to be unusable in that form and required some data engineering tools to make it usable. This is how techniques such as transfer learning came to be applied to fill the gaps created by missing values. However, this posed its own challenges as well as it was not easy to obtain data that would be used for transfer learning. The data had to be collated from aquifers similar to the RTA that resided in different sources.



## CHAPTER 8: CONCLUSIONS

---

### *Engagement with stakeholders*

Various stakeholders (water experts) were engaged to understand their perceptions about AI, big data analytics and the importance of citizen science as a complement to these. The water experts showed awareness of big data analytics and AI aspects and the potential impacts thereof in the water sector. While they generally expressed support for these tools, they emphasised that the human aspect and subject matter should not be lost in the process. They also supported the use of citizen science and its introduction into the mainstream datasets as long as this data was verified.

Other achievements in this task include the acceptance of a proposal to introduce a BSc Hons module on AI at the Wits School of Chemistry. This will go a long way in introducing the students in this field to AI aspects and to increase the visibility of this project as a whole as some examples to be used in the course will come from the work being reported. Rand Water also expressed interest in AI aspects for application in their processes and we have started working with them on some projects related to that.

### *Perceptions of communities regarding citizen science data*

The communities in the Mahikeng and Zeerust areas generally felt that citizen science is important. The majority of the respondents indicated that they depended on borehole water for their livelihood and that of their livestock. A majority indicated that the water was useable in their households with some indicating that they pre-treat their drinking water by boiling and adding bleach. Only 4 out of 60 respondents said that they use bottled water for drinking purposes. From the general descriptions of the water, it was clear that it was hard water.

### *Confirmatory survey*

Our own confirmatory survey including water sampling, analysis, hydrochemical modelling and our “citizen” assessment of the water corroborated the community citizen science and the analytical data in the datasets. We observed that the water was hard, but still useable for households. The water hardness was observable from the scaling on pipes, tanks and toilet systems. The drying of the skin after washing was also quite apparent. Clarity still remained to be made regarding mixtures of groundwater and surface water which were said to be supplied to the households at times. The ratio of the mixtures is not known and the treatment processes involved were not clarified.

### *Big data analytics*

A dataset for the RTA from the IGRAC website was used to perform the big data analytics on the water data. This dataset still had a lot of missing cases. Another dataset from NMMM that covers an area outside the RTA, but with similar characteristics was used for the training and transfer learning conducted to predict the missing cases in the IGRAC dataset. Clustering techniques based on k-means and SOM were used to establish data clusters. Hydrochemical modelling was also applied to assess further water quality aspects. The water was found to be hard as indicated by the communities’ citizen science and our confirmatory survey. The influence

of the dolomitic composition of the aquifer was quite apparent from the correlation conducted. This was further corroborated by the abundance of outcrops of this rock especially in Mahikeng.

The agreement in the three approaches used to conduct the study shows that it is possible to obtain more meaningful results when these are used in tandem.

*Text mining using citizen science and deep learning*

A dataset consisting of the IGRAC data; RTA data (predicted using transfer learning of data from NMMM); citizen science and confirmatory surveys conducted was used. Text mining was applied to establish relationships between textual data (interviews and comments) and numerical data. It was possible to extract the comments and relate them to the categories of borehole water quality (based on alkalinity values). The comments were also rated according to: positive, neutral and negative which translated to good, neutral and bad water quality, respectively.

The similarity model showed that a higher percentage of the boreholes would be rated as positive while those rated negative were fewer. This, however would not necessarily mean that the water quality was good as a sizeable percentage of comments rated as neutral.

The agreement in the approaches used (textual and numerical data) showed the success of building citizen science into conventional big data.

## CHAPTER 9: PROJECT RECOMMENDATIONS

---

The study has shown the levels of awareness about AI, big data analytics and citizen science amongst water experts and communities in general. The levels of trust, especially in ordinary communities remain low thus stifling their contribution in this regard. This has been caused by suspicions that political power could take advantage and manipulate the process of engagement. Most water experts indicated the need to incorporate citizen science into the mainstream datasets.

Collation of data has been a challenge as a result of disparate sources that lack coherence. Where data has been collected, there have been a lot of missing values. Confirmatory surveys assisted in gauging the reliability of citizen science and analytical water data.

Transfer learning proved useful as a deep learning tool to approximate missing values in incomplete datasets. Text mining was used to transform textural data to numeric data and this could be used with hydrochemical data and general big data.

Based on the above, some gaps were identified and the following recommendations are suggested for follow-up research to improve on future findings:

- AI and big data analytics are areas that organisations, including those dealing with water and environmental issues, can harness with the aim of improving water quality and management. They offer the possibility of conducting predictions that can be important for water sustainability.
- Citizen science usage by water utility companies and municipalities has potential to improve the quality of their datasets by extracting extra value that would otherwise be latent. This is incumbent on the data being verified, an aspect that should be built into the questionnaires and data collection tools.
- Communities are willing to be involved in determining water issues that impact on them through participation using citizen science. This implies that there are opportunities for politicians and water authorities to build trust that can see effective participation of communities. This way, the communities will feel empowered to participate positively in service delivery (in this case, water supply) activities.
- Confirmatory surveys can be used as a way to calibrate both citizen science and analytical water data. These can be conducted in small studies that give an overview of the area.
- Where datasets are inadequate and replete with missing data, transfer learning can be used to address this. The only requirement in that case is that the data used for transfer learning should be similar to that in the incomplete dataset.
- The incorporation of citizen science into big data analytics is possible with the use of text mining based on deep learning. This way, questionnaire responses can be processed and classified into appropriate information in numerical form. Thus, comments from communities can be captured through methods such as cellphone communication and classified accordingly. This can be done in real time and not just in batch form, making it possible for communities to monitor and supply constant information about the water they are receiving.

## REFERENCES

---

- AKBAR A, KHAN A, CARREZ F and MOESSNER K (2017) Predictive Analytics for Complex IoT Data Streams. *Internet of Things Journal* **4** (5) 1571-1582.
- ALLEY W, REILLY T and FRANKE O (1999) Sustainability of ground-water resources. Circular 1186. Denver, Colorado: U.S. Geological Survey. <http://dx.doi.org/10.3133/cir1186>. ISBN 978-0-607-93040-5.
- ALTCHENKO Y, LEFORE N, VILLHOLTH K, EBRAHIM G, GENCO A, PIERCE K, WOOLF R, BOTHEPHA BT, MOYO T, KENABATHO P and NIJSTEN GJ (2016) Resilience in the Limpopo Basin: The potential role of the Transboundary Ramotswa Aquifer, baseline report.
- ANURAGA TS, RUIZ L, MOHAN KMS, SEKHAR M and LEIJNSE A (2006) Estimating groundwater recharge using land use and soil data: a case study in South India. *Agricultural Water Management* **84** (2) 65-76. <http://dx.doi.org/10.1016/j.agwat.2006.01.017>.
- BELKHIRI L and MOUNI L (2012) Hydrochemical analysis and evaluation of groundwater quality in El Eulma area, Algeria. *Applied Water Science* **2** (2) 127-133. <https://doi.org/10.1007/s13201-012-0033-6>.
- BERNARD T, NOLAN, MICHAEL N, FIENEN, DAVID L and LORENZ (2015) A statistical learning framework for groundwater nitrate models of the Central Valley, California, USA. *Journal of Hydrology* 902-911. <http://dx.doi.org/10.1016/j.jhydrol.2015.10.025>.
- BEZUIDENHOUT CC, VOS E and TIEDT L (2011) A scoping study on the environmental water groundwater and surface water quality and management in the North-West Province, South Africa. WRC Report No. KV 278/11. Water Research Commission, Pretoria.
- BOTSWANA BUREAU OF STANDARDS (BOS), (2000) Drinking water standard Specification. BOBS Gaborone. Second edition.
- BRAUNE E, CHRISTELIS G and KENABATHO P (2013) Preliminary Evaluation of Kalahari-Karoo aquifer condition; Groundwater Resources Governance in Transboundary Aquifers. URL: <http://grpoundwaterportal.net/sites/default/files/Preliminary> (Accessed: 1 June 2019).
- BRUCE C, METZLER D and TREVOR S (2009) Search engines: Information retrieval in practice. Addison-Wesley, Boston, MA. 136pp.
- BUSCEMA PM, MASSINI G, BREDI M, LODWICK WA, NEWMAN F and ASADI-ZEYDABADI M (2018) Artificial neural networks, studies in Systems, decision and control. Academic Press, Netherlands. 11 pp.
- CRAFFORD JG, HASSAN RM, KING NA, DAMON MC, DEWIT MP, BEKKER S, RAPHOLO BM and OLBRICH BW (2004) An analysis on the social, economic, and environmental direct and indirect costs and benefits of water use in irrigated agriculture and forestry. WRC Report No. 1048/1/04. Water Research Commission, Pretoria.
- DALE R, MOISI H, and SOMERS H (2000) Handbook of natural language processing. CRC Press, New York. 32pp.
- DAVIES J, ROBINS NS, FARR J, SORENSEN J, BEETLESTONE P and COBBING JE (2013) Identifying transboundary aquifers in need of international resource management in the Southern African development community region. *Hydrogeology Journal* **21** (2) 321-330.
- DELEN D and CROSSLAND M (2008) Seeding the survey and analysis of research literature with text mining. *Expert Systems with Applications* **34** (3) 1707-1720.

- DÖRRE J, GERST P and SEIFFERT R (1999) Text mining: Finding nuggets in mountains of textual data. In: Proceedings of the 5th ACM International Conference on Knowledge Discovery and Data Mining. San Diego-ACM Press, 3-4 January 1999, New York.
- FUKUHARA T (2005) Analyzing concerns of people using Weblog articles and real-world temporal data. In: Proceedings of the 2<sup>nd</sup> Annual workshop on the weblogging ecosystem: Aggregation-Analysis and Dynamics, 10 May 2005, Chiba, Japan.
- HADLEY W (2011) The Split-Apply-Combine strategy for data analysis. *Journal of Statistical Software*, **40**(1) 1-29. [http:// www.jstatsoft.org/v40/i01/](http://www.jstatsoft.org/v40/i01/).
- HAMERLY G and ELKAN C (2003) Learning the K in K-Means. In: Seventeenth annual conference on neural information processing systems (NIPS), 281-288 December 2009, MIT Press Cambridge, MA, USA.
- HARROWER MA and BREWER CA (2003) ColorBrewer.org: An online tool for selecting color schemes for maps. *The Cartographic Journal*, **40**(1) 27-37.
- HASTIE T, TIBSHIRANI R and FRIEDMAN J (2009) The elements of statistical learning: Data mining, inference and prediction, 2<sup>nd</sup> ed. Springer Series in Statistics. 34 pp.
- HILL T and LEWICKI P (2007) Statistics methods and applications. Electronic statistics textbook. StatSoft.URL: <http://www.statsoft.com/textbook> (accessed 17 February 2020).
- HILL T, LEWICKI P and QAZAZ C (2007) Multivariate quality control. *Quality Magazine*, **4**(3) 38-45.
- HOLM E (2011) Insurance study sees widespread fraud in NYC. *Wall Street Journal*.URL: <http://blogs.wsj.com/metropolis> (Accessed 17 March 2020).
- HORGAN JM (2012) Review: Computational Statistics and programming in R. *WIREs Computer Statistics* **4** (1) 75-84. [http://dx.doi.org/ 10.1002/wics.183](http://dx.doi.org/10.1002/wics.183).
- HOTH O, STAAB S and STUMME G (2003) Text clustering based on background knowledge. Institute of Applied Informatics and Formal Descriptive Methods, University of Karlsruhe, Germany. 35pp.
- HU M and LIU B (2004) Mining and summarizing customer reviews. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2004), 22-25 August 2004, Seattle, Washington.
- INTERNATIONAL GROUNDWATER RESOURCES ASSESSMENT CENTER (IGRAC) (2012) Transboundary aquifers of the world. IGRAC, International Association of Hydrogeologists, IAH. URL: <https://www.un-igrac.org/tbamap> (Accessed 17 June 2019).
- JAIN AK (2010) Data clustering: 50 years beyond K-means. *Pattern Recognition Letters* **31** (8) 651-666. [http://dx.doi.org/ 10.1016/j.patrec.2009.09.011](http://dx.doi.org/10.1016/j.patrec.2009.09.011).
- KNOLL L, BREUER L and BACH M (2019) Large scale prediction of groundwater nitrate concentrations from spatial data using machine learning. *Science of the Total Environment* **668** (3) 1317-1327. <http://dx.doi.org/10.1016/j.scitotenv.2019.03.045>.
- KWENAMORE KM (2006) Investigation of the levels of antibiotic resistant Enterococcus s and coliform bacteria from groundwater sources within Disobotla and Molopo districts of the North-West Province, South-Africa. MSc dissertation, North-West University, Mafikeng.
- LIN Q, YANG W, ZHENG C, LU K, ZHENG Z, WANG J and ZHU, J. (2018) Deep learning-based approach for forecast of water quality in intensive shrimp ponds. *Indian Journal of Fisheries* **65** (4) 75-80. <https://doi.org/10.21077/ijf.2018.65.4.72559-09>.
- LIU B, HU M and CHENG J (2005) Opinion Observer: Analyzing and comparing opinions on the web.

- In: Proceedings of the 14th International World Wide Web conference (WWW-2005), 10-14 May 2005, Chiba, Japan.
- LOCHBAUM KE, GROSZ BJ and SIDNER CL (2000) Discourse structure and intention recognition. In Dale, R., Moisl, H., & Somers, H. (Eds.), Handbook of Natural Language Processing, Marcel Dekker, Inc., New York. 123-146 pp.
- LOHR S (2009) A \$1 Million Research Bargain for Netflix, and Maybe a Model for Others. New York Times, **21**(5) 45-52.
- MAHGOUB H, RÖSNER D, ISMAIL N and TORKEY F (2008). A text mining technique using association rules extraction. International Journal of Computational Intelligence, **4**(1) 21-28.
- MANNING, CHRIS and HINRICH S (1999) Foundations of statistical natural language processing. Cambridge, MA: MIT Press, Cambridge, MA. 45pp
- MANNING, CHRISTOPHER D, PRABHAKAR R and HINRICH S (2008) Introduction to information retrieval, Cambridge, University Press, MA. 96pp.
- MCGILL BM, ALTCHENKO Y, HAMILTON SK, KENABATHO PK, SYLVESTER SR and VILLHOLTH KG (2019) Complex interactions between climate change, sanitation, and groundwater quality: a case study from Ramotswa, Botswana. Hydrogeology Journal **27** 997-1015. <http://dx.doi.org/10.1007/s10040-018-1901-4>.
- MEYER R (2014) Hydrogeology of Groundwater Region 10: The Karst Belt. WRC Report No. TT 553/14. Water Research Commission, Pretoria.
- MULAMATTATHIL SG, ESTERHUYSEN HA and PRETORIUS PJ (2000) Antibiotic resistant Gram-negative bacteria in a virtually closed water reticulation system. Journal of Applied Microbiology **88** 930-937. <https://doi.org/10.1046/j.1365-2672.2000.01052.x>.
- NIJSTEN GJ, CHRISTELIS G, VILLHOLTH KG, BRAUNE E and GAYE CB (2018) Transboundary aquifers of Africa: Review of the current state of knowledge and progress towards sustainable development and management. *Journal of Hydrology* **20** 21-34. <https://doi.org/10.1016/j.ejrh.2018.03.004>.
- NISBET R, ELDER J and MINER G (2009) Handbook of statistical analysis and data mining Applications. Elsevier, Burlington, MA. 123pp.
- OWEN R (2011) Groundwater needs Assessment Limpopo Basin Commission LIMCOM. Waternet Report No. TT 123/11. Africa Groundwater Network.
- PIETERSEN K, BEEKMAN HE and HOLLAND M (2011) South African groundwater governance case study. WRC Report No. KV 273/11. Water Research Commission, Pretoria.
- PIETERSEN K, BEEKMAN HE and HOLLAND M (2011) South African groundwater governance case study. WRC Report No. KV 273/11. Water Research Commission, Pretoria.
- POLON J (2011) Text mining case study: Text mining for health insurance. In: Proceedings at the Society of Actuaries Health Meeting, 3-6 July 2011, Boston, MA.
- RANGANAI TR, GOTLOP-BOGATSU Y and MAPHANYANE J (2001) Hydrochemical and Geophysical evaluation of groundwater pollution in the Ramotswa wellfield, SE Botswana. BIE2001-Technical Papers 193-200.
- RATINOV L and ROTH D (2009) Design challenges and misconceptions in named entity recognition. In: Proceedings of the Annual Conference on Computational Natural Language Learning (CoNLL). 147-155 June 2009, Colorado.



- RENDERS J (2004) Kernel methods for natural language. In: *Processing Learning Methods for Text Understanding and Mining*, 26-29 January 2004, Grenoble, France.
- SANCHEZ-MARTOS F, AGUILERA PA, GARRIDO-FRENICH A, TORRES JA and PULIDO-BOSCH A (2002) Assessment of groundwater quality by means of self-organizing maps: Application in a semiarid area. *Environmental Management* **30** (5) 716-726. <https://doi.org/10.1007/s00267-002-2746-z>.
- SANCHEZ-MARTOS F, AGUILERA PA, GARRIDO-FRENICH A, TORRES JA and PULIDO-BOSCH A (2002) Assessment of groundwater quality by means of self-organizing maps: Application in a semiarid area. *Environmental Management* **30** (5) 716-726. <https://doi.org/10.1007/s00267-002-2746-z>.
- SENI G and ELDER J (2010) *Ensemble methods in data mining: Improving accuracy through combining predictions*. Chicago, Morgan and Claypool Publishers, **4**(3) 45-51.
- SHUKUR OB and LEE MH (2015) Imputation of missing values in daily wind speed data using hybrid AR-ANN method. *Modern Applied Science* **9** (11) 1-11. <https://doi.org/10.5539/mas.v9n11p1>.
- SMILEY D and PUGH E (2009) *Solr 1.4: Enterprise Search Server*. Packt Publishing, Birmingham, England, UK. 65pp.
- SMITH A and HUMPHREYS M (2006) Evaluation of unsupervised semantic mapping of natural language with Leximancer concept mapping. *Behavior Research Methods*, **38**(2) 262.
- SMITH LI (2002) A tutorial on Principal Components Analysis Introduction In: John Wiley & Sons (ed.). Academic Press, New York.
- SOUMEN C (2002) *Mining the web: Analysis of hypertext and semi-structured data*. Morgan Kaufmann, San Francisco. 56pp.
- STADLER S, TALMA AS, TREDoux G and WRABEL J (2012) Identification of sources and infiltration regimes of nitrate in the semi-arid Kalahari: Regional differences and implications for groundwater management. *Water SA* **38** (2) 213-224. <https://hdl.handle.net/10520/EJC120097>.
- STAUDT M (2003) *Production of Environmental Hydrogeology Maps Using GIS for the Ramotswa Project Area, South East District, Botswana*. Department of Geological Survey (DGS), Lobatse, Botswana.
- STAUDT MH AND VOGEL (2004) Environmental hydrogeology of the dolomite aquifer in Ramotswa, Botswana. In: Stephenson, D., Shemang, E.M.T.R. Chaoka (Eds.): *Water Resources of Arid Areas. Proceedings of the International Conference on Water Resources of Arid and Semi-Arid Regions of Africa (WRASRA)*, Gaborone, Botswana, A.A. 371-377 September 2004, Balkema Publishers.
- SUREKA A, DE S and VARMA K (2008). *Mining automotive warranty claims data for effective root cause analysis*. Lecture Notes in Computer Science, Indiana University.
- TALMA A and VAN W (2013) *Rainfall and groundwater isotope Atlas*. In: Abiye (ed). *The use of isotope hydrology to characterize and assess water resources in South(ern) Africa*. WRC Report No TT 570/13. Water Research Commission, Pretoria.
- VIDHYA KA and AGHILA G (2010) Text mining process, techniques and tools: an Overview. *International Journal of Information Technology and Knowledge Management* **2**(2), 613-622.
- VILLHOLTH KG, TØTTRUP C, STENDEL M and MAHERRY A (2013) Integrated mapping of groundwater drought risk in the Southern African Development Community (SADC) region. *Hydrogeology Journal* **21** (4) 863-885.
- VOGEL H, MOKOKWE K and SETLOBOKO T (2004) *Nitrate hotspots and salinity levels in groundwater in the Central District of Botswana*. Environmental Geology Division Report No 00-12-04. Department of

Geological Survey, Lobatse, Botswana.

WAKIDA FT and LERNER DN (2005) Non-agricultural sources of groundwater nitrate: a review and case study. *Water Research* **39** (1) 3-16.

WANG B, STRELAKOS PM and JOKELA B (2000) Nitrate source indicators in ground water of the Scimitar Subdivision, Peters Creek area. In: Anchorage (ed) Alaska US department of the Interior and Geological Survey. *Water-Resources Investigations Report* 00-4137.

WATER RESEARCH COMMISSION (WRC) (2003) Understanding Groundwater: A stakeholder's guide to the North West dolomite aquifer.

WEAVER JM, CAVE L and TALMA AS (2007) Groundwater Sampling. WRC Report No. TT 303/07. Water Research Commission, Pretoria.

WORLD HEALTH ORGANISATION (WHO) (2004) Guidelines for Drinking-Water Quality. Background document for development of WHO Guidelines for Drinking-water Quality. Originally published in Guidelines for drinking-water quality, 3<sup>rd</sup> ed. World Health Organization, Geneva.

WORLD HEALTH ORGANISATION (WHO) (2011) Total dissolved solids in Drinking-water. Background document for development of WHO Guidelines for Drinking-water Quality. Originally published in Guidelines for drinking-water quality, 12nd ed. Vol. 2. Health criteria and other supporting information. World Health Organization, Geneva.

WORLD HEALTH ORGANISATION (WHO) (2015) Total dissolved solids in Drinking-water. Background document for development of WHO Guidelines for Drinking-water Quality. Originally published in Guidelines for drinking-water quality, 2nd ed. Vol. 2. Health criteria and other supporting information. World Health Organization, Geneva.

WORMALD R, ECKARDT F and VEARNCOMBE JS (2003) Spatial distribution analysis of pans in Botswana: The importance of structural control. *South African Journal of Geology* **106** 287-290.

ZHENG Q, MA T, WANG Y, YAN Y, LIU L and LIU L (2017) Hydrochemical characteristics and quality assessment of shallow groundwater in Xincai river basin, Northern China. *Procedia Earth and Planetary Science* **17** 368-371. <https://doi.org/10.1016/j.proeps.2016.12.093>.

## APPENDIX 1: WATER EXPERTS SURVEY

---

### INTRODUCTION AND CONSENT

Hello. My name is \_\_\_\_\_. I am from the University of the Witwatersrand in Johannesburg]. We are conducting a survey about the application of big data in the water domain, and specifically about experts' opinions on Citizen Science data in water resources management. The information we collect will be used for academic research only. The questions usually take about 15 to 20 minutes. All of the answers you give will be confidential and will not be shared with anyone other than members of our survey team. You don't have to be in the survey, but we hope you will agree to answer the questions because your views are important. If I ask you any question you do not want to answer, just let me know and I will go on to the next question or you can stop the interview at any time. In case you need more information about the survey, you may contact the person listed on this card.

### (GIVE CARD WITH CONTACT INFORMATION)

Before we begin, please answer the following questions by ticking "Yes" or "No"

Do you understand what I explained to you?	Yes	No
Do you understand that we can stop at any time?	Yes	No
Do you have any questions that you want to ask me before we start?	Yes	No
Can I start asking you the questions on the form?	Yes	No

I have explained the project and the implications of being interviewed to the respondent, and I believe that the consent is informed and that he/she understands the implications of participation.

Interviewer's name \_\_\_\_\_

Signature of interviewer: \_\_\_\_\_

Date: \_\_\_\_\_

IDENTIFICATION		
Expertise, e.g. water engineer, data analyst		
Work location / industry		
Interviewee details	What is your age category?	
	18-24 years	
	25-35 years	
	36-50 years	
	50+ years	
	Gender	
	Male	
	Female	
	Education level	
	Bachelor degree	
	Honours degree	
	Master degree	
	PhD degree	

## **Big Data**

### **General understanding of Big Data**

*(These are questions to gauge the “world view” of the participant and their understanding / perception of their role/ experience with Big Data.)*

- a) What does Big Data (and analytics) mean, in your opinion?
- b) Do you think Big Data Analytics would be useful for the management of water resources and water quality?
- c) Do you have any reservations/uncertainty regarding aspects related to big data such as artificial intelligence and robotics?

## **Citizen Science**

### **1) General understanding of Citizen Science**

*(These are questions to gauge the “world view” of the participant and their understanding / perception of their role/ experience with Citizen Science.)*

- d) What does Citizen Science mean, in your opinion?
- e) Do you think that other stakeholders that use the data or participate in Citizen Science might think of it differently?
- f) Please tell me a little bit about your experience and direct/indirect involvement with of Citizen Science?

### **2) General perceptions of Citizen Science**

*(These questions are for gaining insight into the different ways the participant views the benefits (or limitations) of various aspects of citizen science.)*

- a) Have you seen / *Do you believe* Citizen Science programs or data be of benefit for water resources management?
- *What do you think makes it beneficial?*
  - *Do you think there are any barriers or limitations to creating benefit?*
  - *How do you think these barriers could be overcome?*
- b) What are your thoughts on using Citizen Science as a tool for:
- *increasing public scientific literacy*
  - *environmental stewardship*
  - *environmental democracy*
- c) What are your thoughts on using Citizen Science data for addressing missing data in data sets?
- d) What is your perception on the value of Indigenous-knowledge based citizen science?
- e) Do you perceive any negative impacts of Citizen Science programs or data?

### **3) Perceptions on Citizen Science methods and data quality**

*(These questions attempt to decipher attitudes around limitations of citizen science methods and data quality.)*

- a) What value do you think that Citizen Science water quality data can add to conventional data and our understanding or prediction of water quality?
- b) Do you think there are any barriers in achieving high-quality, highly-usable Citizen Science data?
- c) In your opinion, what would a useable set of Citizen Science data comprise?

- d) What do you think it would take for Citizen Science data to be widely accepted by big data analysts?
  
- e) Would you be / are you willing to integrate citizen science data with conventional water quality data in your role as a water professional?



## APPENDIX 2: CITIZEN SCIENCE DATA SURVEY

---

### Community Survey

IDENTIFICATION		
Place name		
Household number		
Interviewee details	Age	
	Sex	
	Education level	
	None	
	Primary	
	Secondary	
	Vocational	
	University degree	
	Employment status	
	Employed	
	Not employed	
Date		
Interviewer's name		
Language of interview		
Native language of respondent		

---

### INTRODUCTION AND CONSENT

Hello. My name is \_\_\_\_\_. I am from the University of the Witwatersrand in Johannesburg]. We are conducting a survey about water in the Zeerust and Mahikeng area. The information we collect will be used for academic research only. Your household was selected for the survey. I would like to ask you some questions about your water use. The questions usually take about 15 to 20 minutes. All of the answers you give will be confidential and will not be shared with anyone other than members of our survey team. You don't have to be in the survey, but we hope you will agree to answer the questions since your views are important. If I ask you any question you don't want to answer, just let me know and I will go on to the next question or you can stop the interview at any time. In case you need more information about the survey, you may contact the person listed on this card.

*(GIVE CARD WITH CONTACT INFORMATION)*

Before we begin, please answer the following questions by ticking "Yes" or "No"

Do you understand what I explained to you?	Yes	No
--------------------------------------------	-----	----

Do you understand that we can stop at any time?	Yes	No
Do you have any questions that you want to ask me before we start?	Yes	No
Can I start asking you the questions on the form?	Yes	No

I have explained the project and the implications of being interviewed to the respondent, and I believe that the consent is informed and that he/she understands the implications of participation.

Name of interviewer: \_\_\_\_\_

Signature of interviewer: \_\_\_\_\_

Date: \_\_\_\_\_

### Survey Instrument

1. On a scale of 1 ( <b>Strongly disagree</b> ) to 6 ( <b>Strongly agree</b> ), please rate your agreement with the following statements						
	1	2	3	4	5	6
a) Most people are basically honest.						
b) Most people are trustworthy.						
c) Most people are basically good and kind.						
d) Most people are trustful of others.						
e) I am trustful.						
f) Most people will respond in kind when they are trusted by others.						
2. On a scale of 1 ( <b>do not trust their information at all</b> ) to 6 ( <b>trust their information very much</b> ), please tell me to what extent you trust the following sources of information to tell you the truth about the quality of water of your area (Zeerust or Mahikeng).						
	1	2	3	4	5	6
a) Workers at the supplying authority's water treatment facility						
b) Scientists working for government agencies						
c) University scientists doing research in water						

3. On a scale of <b>1 (Strongly disagree)</b> to <b>6 (Strongly agree)</b> , please rate your agreement with the following statements						
	<b>1</b>	<b>2</b>	<b>3</b>	<b>4</b>	<b>5</b>	<b>6</b>
a) All things considered, the world is better off because of science and technology.						
b) Scientists know best what is good for the public						
c) Water quality data collected by citizens adds value to data collected by scientists						
d) Water quality data collected by citizens can be useful for decision-making where scientific data is not available.						
e) Scientists should incorporate information collected by citizens, e.g. colour and taste, in determining water quality						
f) I feel informed about the quality of the water I use for daily purposes						

### Descriptive data

No	Question	Category
1	What is the main source of drinking water for members of your household?	<b>1. Piped water</b> <ul style="list-style-type: none"> <li>(i). Piped into dwelling</li> <li>(ii). Piped to yard/plot</li> <li>(iii). Piped to neighbor</li> <li>(iv). Public tap/standpipe</li> <li>(v). Tube well or borehole</li> </ul> <b>2. Dug well</b> <ul style="list-style-type: none"> <li>(i). Protected well</li> <li>(ii). Unprotected well</li> </ul> <b>3. Water from spring</b> <ul style="list-style-type: none"> <li>(i). Protected spring</li> <li>(ii). Unprotected spring.</li> </ul> <b>4. Rainwater.</b> <b>5. Tanker truck</b> <b>6. Cart with small tank</b> <b>7. Surface water (river/dam/ Lake/pond/stream/canal/ Irrigation channel)</b> <b>8. Bottled water</b>

No	Question	Category
2	What is the main source of water used by your household for other purposes such as cooking and hand washing?	<b>1. Piped water</b> (i). Piped into dwelling (ii). Piped to yard/plot (iii). Piped to neighbor (iv). Public tap/standpipe (v). Tube well or borehole <b>2. Dug well</b> (i). Protected well (ii). Unprotected well <b>3. Water from spring</b> (i). Protected spring (ii). Unprotected spring. <b>4. Rainwater</b> <b>5. Tanker truck</b> <b>6. Cart with small tank</b> <b>7. Surface water</b> (river/dam/ Lake/pond/stream/canal/ Irrigation channel)
3	Where is that water source located?	<b>1.</b> In own dwelling _____ <b>2.</b> In own yard/plot _____ <b>3.</b> Elsewhere _____
4	How long does it take to go there, get water, and come back?	<b>1.</b> Minutes _____ <b>2.</b> I don't know _____ <b>3.</b> Not applicable _____
5	In the past two weeks, was the water from this source not available for at least one full day?	<b>1.</b> Yes _____ <b>2.</b> No _____ <b>3.</b> Don't know _____
6	Do you consider the water you drink from the named source to be safe for consumption?	<b>1.</b> Yes _____ <b>2.</b> No _____ <b>3.</b> Don't know _____
7	How do you check / confirm if the water is safe? <i>RECORD ALL MENTIONED</i>	<b>1.</b> Visual characteristics, e.g. colour, clarity <b>2.</b> Taste <b>3.</b> Other
9	What do you usually do to make the water safer to drink?  Anything else?	<b>1.</b> Boil <b>2.</b> Add bleach/chlorine <b>3.</b> Strain through a cloth <b>4.</b> Use water filter (ceramic/ Sand/composite/

No	Question	Category					
	<i>RECORD ALL MENTIONED</i>	<b>5.</b> Solar disinfection <b>6.</b> Let it stand and settle <b>7.</b> Other <b>8.</b> Don't know					
10	Please describe to me, verbally or using a drawing, how you carry out this process of making the water safe.	(WRITE DESCRIPTION OR PROVIDE MATERIALS FOR DRAWINGS. ENCOURAGE RESPONDENTS TO ANNOTATE DRAWINGS).					
11	Do you trust the information supplied by water authorities about groundwater / piped water?	<b>1.</b> Yes <b>2.</b> No					
12.	How willing would you be to share your assessments of water quality and safety with water treatment and supplying authorities?	1(Not at all willing)	2	3	4	5	6 (very willing)

## PPENDIX 3: SUPPLEMENTARY DATA - R CODES USED

---

### 1. Transfer learning

```
library(keras)
install_keras()
library(mlbench)
library(dplyr)
library(magrittr)
library(neuralnet)
data = datatr[,c("F-", "EC", "TDS", "NO3", "pH", "SO42-", "Cl-", "CaCO3", "Ca2+", "Mg2+", "K+", "Na+")]
str(data)
data %<>%mutate_if(is.double, as.numeric)
n<-neuralnet(Na~K+Mg+Ca+Cl+SO4+pH+EC+NO3+FI+TDS+CaCO3,
data = data, hidden = c (10,5),
            linear.output = F,
            lifesign = 'full',rep=1)
plot(n, col.hidden = 'darkgreen',
     col.hidden.synapse = 'darkgreen',
     show.weights = F,information = F,
     fill = 'lightblue')
data <-as.matrix(data)
dimnames(data)<-NULL
set.seed (123)
ind<-sample(2,nrow(data),replace = T, prob= c(7,3))
training <-data[ind==1,1:11]
test<-data[ind==2,1:11]
trainingtarget<-data[ind==1,12]
testtarget<-data[ind==2,12]
y=trainingtarget
m<-colMeans(training)
s<-apply(training,2,sd)
training <-scale(training, center = m, scale = s)
test <-scale(test, center = m, scale = s)
```

#### ✓ Create Model

```
model <-keras_model_sequential()
model%>%
  layer_dense(units =200, activation = 'relu', input_shape = c(11))%>%
  layer_dense(units = 1)
```

#### ✓ Compile model

```
model %>% compile(loss = 'mse', optimizer = 'rmsprop', metrics = 'mae')
mymodel <-model %>% fit(training,trainingtarget,epoch = 100,batch_size = 32,
  validation_split = 0.2)
model%>%evaluate(test,testtarget)
pred <-model%>%predict(test)
mean((testtarget-pred)^2)
plot(testtarget, pred)
```



```
cor(testtarget, pred)
✓ Fine-tune model
set.seed(1234)
model <- keras_model_sequential()
model %>% layer_dense(units = 500, activation = 'relu', input_shape = c(11)) %>%
  layer_dropout(rate = 0.3) %>% layer_dense(units = 280, activation = 'relu') %>%
  layer_dropout(rate = 0.2) %>% layer_dense(units = 150, activation = 'relu') %>%
  layer_dropout(rate = 0.1) %>% layer_dense(units = 1)
summary(model)
✓ Compile model
model %>% compile(loss = 'mse', optimizer = optimizer_rmsprop(lr = 0.001), metrics = 'mae')
✓ Fit model
mymodel <- model %>% fit(training, trainingtarget, epoch = 100, batch_size = 32, validation_split = 0.2)
✓ Evaluate
model %>% evaluate(test, testtarget)
pred <- model %>% predict(test)
mean((testtarget - pred)^2)
plot(testtarget, pred)
cor(testtarget, pred)
✓ Transfer learning
data2 = as.matrix(RIMS_prediction[, c("FI", "EC", "TDS", "NO3", "pH", "SO4", "CL", "CACO3", "Ca", "Mg", "K")])
dimnames(data2) <- NULL
test2 = data2
test2 <- scale(test2, center = m, scale = s)
model %>% evaluate(test, testtarget)
pred2 <- model %>% predict(test2)
mean((testtarget - pred)^2)
plot(testtarget, pred)
```

## 2. Self-organizing map

```
library(cluster)
library(NbClust)
library(kohonen)
library(ggplot2)
library(gridExtra)
library(scales)
library(kohonen)
library(RColorBrewer)
library(fields)
data = Hydrochemical_assessment[, -(1:4)]
df = data[, 1:4]
df = data2[, 2:5]
data2 = Hydrochemical_assessment[, -(2:4)]
dt <- scale(df, center = TRUE)
str(data)
is.na(dt)
library(corrplot)
corrplot(cor(data))
# Prepare SOM
set.seed(590507)
```

```

library(kohonen)
som1 <- som(dt,
  somgrid(6,6, "hexagonal"),
  rlen=500,
  keep.data=TRUE)
myPal1=colorRampPalette(c("black","orange","red","green"))

plot(som1, type="codes", palette.name = myPal1,
  main="Codes", shape="straight",
  border ="gray", heatkey = TRUE)
plot(som1, type = "counts", palette.name =myPal1, main="Codes",
  shape="straight",
  border ="gray")
cds <- as.data.frame(som1$codes)
wss <- (nrow(cds)-1)*sum(apply(cds,2,var))
for (i in 2:6){
  wss[i] <- sum(kmeans(cds,centers=i)$withinss)
}

par(mar = c(8,5,8,2))
plot(1:6, wss, type="b",
  xlab="Number of Clusters",
  ylab="Within groups sum of squares",
  main="Within cluster sum of squares (WCSS)",
  col="blue",
  lwd =2)

nCls =3
som1.km <- kmeans(cds, nCls, nstart = 20)
str(som1.km)
som1.km$centers
som1.km$totss
som1.km$cluster
MyPal3 <- c("grey80", 'aquamarine', 'burlywood1')
par(mar = c(0,5,0,2))

plot(som1, type="codes",
  palette.name= myPal1,
  bgcol = MyPal3[som1.km$cluster],
  main = "SOM-k-mean cluster",
  shape="straight",
  border ="red")
legend("right", x=8,y=4, cex=1.5,
  title="Cluster", legend = c(1:nCls),
  fill= MyPal3[c(1:nCls)])

SOM.clss <- as.data.frame(som1$unit.classif)
names(SOM.clss) <- "Cell.Nmbr"
unique(SOM.clss)
kMns.clst <- as.data.frame(som1.km$cluster)
names(kMns.clst) <- "Clstr"
kMns.clst$Cell.Nmbr <- 1:nrow(kMns.clst)

```

```
dt.clst <- merge(SOM.clss,kMns.clst,by="Cell.Nmbr")
data.clst <- cbind(data,dt.clst)
aggregate(data.clst[,1:4],
          by=list(data.clst$Clstr),
          FUN=mean)

km <- kmeans(subset(data, select=-c(WQI)), centers=4)
km3<-kmeans(subset(data,select=-c(WQI,SAR)),centers=3)
clusplot(subset(data, select=-c(WQI)), km$cluster, main="cluster 4 : All variables")
clusplot(subset(data, select=-c(WQI,SAR)), km3$cluster, main="cluster 3 : cluster")

data$cluster <- as.factor(km$cluster)
qplot(WQI,TH, colour=cluster, data=data)
data$cluster3 <- as.factor(km3$cluster)
qplot(data$`Na%`,TH,colour=cluster3,data=data)

p1 <-qplot(WQI, fill=cluster, alpha=.5, data=data, geom="density") +
scale_alpha(guide="none")
p2 <-qplot(SAR, fill=cluster, alpha=.5, data=data, geom="density") +
theme(legend.position="none")
p3 <-qplot(data$`Na%`, fill=cluster, alpha=.5, data=data, geom="density") +
theme(legend.position="none")
p4 <-qplot(data$TH, fill=cluster, alpha=.5, data=data, geom="density") +
theme(legend.position="none")
grid.arrange(p1, p2, p3, p4, ncol=2, nrow=2)

x <- ggplot(data, aes(x=factor(1), fill=cluster))
x + geom_bar(width=1) + coord_polar(theta="y")

k2m <- kmeans(dt, centers = 3, iter.max = 500, nstart = 10)
k2m
image(t(dt)[, nrow(dt):1], yaxt = "n", main = "Original Data")
image(t(dt)[, order(k2m$cluster)], yaxt = "n", main = "Clustered Data")

# Cluster
library("kohonen")
library("NbClust")
library("factoextra")

# Visualization
library("ggplot2")
library("car") # associated with the book 'An R Companion to Applied Regression', Third Edition, by John Fox
and Sanford Weisberg.
library("rgl")

colors <- c('red','orange','green3','deepskyblue','blue',
            'darkorchid4','violet','pink1','tan3','black')

options(rgl.printRglwidget = TRUE)
scatter3d(x = (data$WQI), y = (data$SAR), z = (data$`Na%`),
          groups = as.factor(k2m$cluster),
          xlab = "Frequency (Log-transformed)", ylab = "Monetary Value (log-
```

```

transformed)",
zlab = "Recency (Log-transformed)",
surface.col = colors, axis.scales = FALSE,
surface = TRUE, # produces the horizontal planes through the graph at each level
of monetary value
fit = "smooth",
# ellipsoid = TRUE, # to graph ellipses uses this command and comment out "surface = TRUE"

grid = TRUE, axis.col = c("black", "black", "black"))

som.rfm <- som(dt, grid = somgrid(20, 20, "hexagonal"))
plot(som.rfm,type = "counts",main = "Som")
require(NbClust)
set.seed(1)
nc <- NbClust(dt, method="kmeans")
fviz_nbclust(dt, kmeans, method = "wss")
fviz_nbclust(dt, kmeans, method = "silhouette")
length(k2m$cluster)
length(som.rfm$unit.classif)
dt$clustersom = as.factor(som.rfm$unit.classif)
dt$clusterk=as.factor(k2m$cluster)
kable(head(dt))
library(reshape2)
dsClusters = melt(data.frame(data))
kable(head(dsClusters))

data.n <- scale(as.matrix(subset(data, select=-c(WQI))))
dt2<-scale(as.matrix(data))
set.seed(1)
som_model <- som(dt2 , grid = somgrid(3, 3, "rectangular"))
str(som_model)
plot(som_model, main = "SOM")
plot(som_model, type="counts", main = "cluster size")
plot(som_model, type="quality", main = "mapping quality")

coolBlueHotRed <- function(n, alpha = 1) {
  rainbow(n, end=4/6, alpha=alpha)[n:1]}
for (i in 1:ncol(som_model$data))
  plot(som_model, type="property",property=som_model$codes[,1],
main=dimnames(som_model$data)[[2]][1], palette.name=coolBlueHotRed)

Hexagon <- function (x, y, unitcell = 1, col = col) {
  polygon(c(x, x, x + unitcell/2, x + unitcell, x + unitcell,
    x + unitcell/2), c(y + unitcell * 0.125,
      y + unitcell * 0.875,
      y + unitcell * 1.125,
      y + unitcell * 0.875,
      y + unitcell * 0.125,
      y - unitcell * 0.125),
    col = col, border=NA)}

somClusters <-kmeans(som_model2$codes,centers=3)

```

```
somClusters
plot(0, 0, type = "n", axes = FALSE, xlim = c(0, som_model2$grid$xdim), ylim = c(0, som_model2$grid$ydim),
xlab = "", ylab = "", asp = 1)
ColRamp <- rev(designer.colors(n=k, col=brewer.pal(k,"Set1")))
ColorCode <-rep("#FFFFFF",length(somClusters$cluster))

for (i in 1:length(somClusters$cluster))
  ColorCode[i] <- ColRamp[somClusters$cluster[i]]

offset <- 0.5
for (row in 1:som_model2$grid$ydim) {
  for (column in 0:(som_model2$grid$xdim-1))
    Hexagon(column + offset, row - 1, col = ColorCode[row + som_model2$grid$ydim * column])
  offset <- ifelse(offset, 0, 0.5)
}
data$NavsCI<-data$`Na+`/data$`Cl-`
data$cluster <- as.factor(km$cluster)
qplot(data$NavsCI,data$EC, colour=cluster, data=data)
data$cluster3 <- as.factor(km3$cluster)
qplot(data$`Na%`,`TH`,colour=cluster3,data=data)
```

### 3. Text mining

#### Model building

##### Trains and testing set

```
# indexes <- createDataPartition(df2$Rate, times = 1, p = 0.7, list = FALSE)
# train <- df2[indexes,]
# test <- df2[-indexes,]
```

##### Converting text to digital

```
# train.tokens[[357]]
[1] "water" "is" "good" "nonsalty" "I" "dont" "use" "to" "much" "soap" "to" "wash"
[13] "clothes" "and" "dishes"
# train.tokens[[357]]
```

*First bag-of words model*

##### Cross validation

```
# cv.folds <- createMultiFolds(train$Rate, k = 10, times = 3)
# cv.cntrl <- trainControl(method = "repeatedcv", number = 10, repeats = 3, index = cv.folds)
# rpart.cv.1 <- train(Rate ~ ., data = train.tokens.df, method = "rpart", trControl = cv.cntrl, tuneLength = 7)
# inverse.doc.freq <- function(col) { corpus.size <- length(col) doc.count <- length(which(col > 0))
log10(corpus.size / doc.count)}
# tf.idf <- function(x, idf) {x*idf}
# train.tokens.df <- apply(train.tokens.matrix, 1, term.frequency)
# train.tokens.idf <- apply (train.tokens.matrix, 2, inverse.doc.freq)
# train.tokens.tfidf <- apply(train.tokens.df, 2, tf.idf, idf = train.tokens.idf)
  The train.tokens.tfidf was transposed to a matrix (The result can be downloaded from the following link (a href="#">\(\))).
# train.irlba <- irlba(t(train.tokens.tfidf), nv = 32, maxit = 200)
# total.time <- Sys.time() - start.time
# total.time
# sigma.inverse <- 1 / train.irlba$d
```

```
# u.transpose <- t(train.irlba$u)
# document <- train.tokens.tfidf[1,]
# document.hat <- sigma.inverse * u.transpose %*% document
# rf.cv.1 <- train(Rate ~ ., data = train.svd, method = "rf", trControl = cv.cntrl, tuneLength = 7)
# confusionMatrix(train$Rate, rf.cv.1$finalModel$predicted)

Predictive Modeling

Tokenization
# test.tokens <- tokens(test$Text, what = "word", remove_numbers = TRUE, remove_punct =
  TRUE, remove_symbols = TRUE, remove_hyphens = TRUE)

Lower case the tokens
# test.tokens <- tokens_tolower(test.tokens)

Stopword removal
# test.tokens <- tokens_select(test.tokens, stopwords(),selection = "remove")

Stemming
# test.tokens <- tokens_wordstem(test.tokens, language = "english")

Add bigrams
# test.tokens <- tokens_ngrams(test.tokens, n = 1:2)

Convert n-grams to quanteda document-term frequency matrix
# test.tokens.dfm <- dfm(test.tokens, tolower = FALSE)
# test.tokens.dfm <- dfm_select(test.tokens.dfm, pattern = train.tokens.dfm, selection = "keep")
# test.tokens.matrix <- as.matrix(test.tokens.dfm)
# test.tokens.df <- apply(test.tokens.matrix, 1, term.frequency)
# str(test.tokens.df)
# test.tokens.tfidf <- apply(test.tokens.df, 2, tf.idf, idf = train.tokens.idf)
# test.tokens.tfidf <- t(test.tokens.tfidf)
# test.svd.raw <- t(sigma.inverse * u.transpose %*% t(test.tokens.tfidf))
# test.svd <- data.frame(Rate = test$Rate, test.svd.raw, TextLength = test$TextLength)
# test.similarities <- rbind(test.svd.raw, train.irlba$v[Positive.indexes,])
# test.similarities <- cosine(t(test.similarities))

Random forest prediction model
#preds <- predict (rf.cv.3, test.svd)
```



