A REVIEW OF THE USE OF R PPROGRAMMING FOR DATA SCIENCE RESEARCH IN BOTSWANA

Simisani Ndaba

Department of Computer Science, University of Botswana, Gaborone, Botswana

ABSTRACT

R is widely used by researchers in the statistics field and academia. In Botswana, it is used in a few research for data analysis. The paper aims to synthesis research conducted in Botswana that has used R programming for data analysis and to demonstrate to data scientists, the R community in Botswana and internationally the gaps and applications in practice in research work using R in the context of Botswana. The paper followed the PRISMA methodology and the articles were taken from information technology databases. The findings show that research conducted in Botswana that use R programming were used in Health Care, Climatology, Conservation and Physical Geography, with R part as the most used R package across the research areas. It was also found that a lot of R packages are used in Health care for genomics, plotting, networking and classification was the common model used across research areas.

KEYWORDS

R Programming, Botswana, R Package, Research Area, Data Analysis

1. INTRODUCTION

R is a programming language which is used for statistical computing, data analysis and graphic production, and can be obtained free of charge from the Internet [1].An advantage of the R ecosystem is the powerful set of add-on packages that can be used to perform a range of tasks from experimental design. R programming language is one of the most popular means of introducing computing into data science, data analytics, and statistics curricula[2]. R is part of The Carpentries[3],a global non-profit organization that teaches practical data science skills to researchers through active learning workshops.

1.1. Background

Research in Botswana is guided by [4]. [5] wrote that while development-oriented research is a priority, in the interest of expanding knowledge in various fields, research of a more academic and theoretical nature is permitted wherever possible. The research topics with the greatest number of publications are in the areas of Geosciences, Multidisciplinary Ecology; Environmental Sciences; Water Resources and Veterinary Sciences. Botswana has a national research, science and technology plan [6], which is informed by the need for a centre of excellence to earn a reputation as a significant resource for the progress of science and technology and the spread of innovation, a strong dependency on imported fuel such as oil and electricity, a resurgence of diseases such as TB and increasing HIV-related infections, a wealth of untapped indigenous knowledge in traditional Botswana society, and importance of information and communication technology as vital for the country's future as a pervasive enabler of industry and developmental solutions [5].The statistical research conducted in Botswana

is descriptive with the common use of STATA, SPSS, and ATLAS-TI used by [7], [8] and [9] for example. R programming is mostly used by statisticians despite its growing increased use in data science, however, very few research conducted in Botswana have used R programming for their data analysis.

1.2. Objective

This review paper aims to provide a concise snapshot of the research to date investigating research in Botswana that has used R programming for their data analysis. The review also captures how research used R programming capabilities in their data analysis work. Previous reviews have demonstrated Python programming to be robust and scalable for data analysis, but no review has comprehensively mapped R programming applications within the context of research in Botswana. Such a review would equip both data scientists and practitioners in the methods and applications of R programming in Botswana. It would also highlight the challenges of using R programming in Botswana, as well as identify gaps in the data science field and potential opportunities for further research.

First, the paper outlines the search strategies used to find relevant literature. Next, the paper conducts a synthesis of the literature, describing both the R programming and research in Botswana of each article. Finally, the paper summarizes the extant research and the implications for future work.

2. METHODOLOGY

2.1. Search Strategy

The paper followed the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses)for reporting in a systematic review. PRISMA was chosen because it is the recognized standard for reporting evidence in systematic reviews and meta-analyses and the standards are endorsed by organizations and journals. The search strategy was adapted from [10] and [11] who followed the PRISMA methodology. As R programming and Botswana fall under different themes, the search was conducted in only information technology databases to access research done in Botswana that have used R programming. A literature search was conducted using the information technology databases IEEE Xplore and the ACM Digital Library were searched. Finally, the database that index both fields, including Web of Science, were searched. Goggle scholar was also used as a source of data collection. The search period for relevant research was conducted in September 2022. The search terms included variations in the term for the following:

• (a) R programming Botswana (*R statistics Botswana*, R programming Botswana*, R statistical analysis Botswana*)

The search was conducted on titles, keywords, and abstracts with *AND* entered into the database search to link different category (a) of search terms. Truncation symbols (∗) were used to search for all possible forms of a search term. Forward reference searching, that is, examining the references cited in these research, and backward reference searching, that is, reviewing the references cited in these research, were applied to identify further research that met the inclusion criteria.Table 1 below shows the criteria that were met to include and exclude articles from the review.

Table 1. Inclusion and Exclusion criteria followed that were followed.

Inclusion Criteria	**Exclusion Criteria**
The article reported on a method or application of R programming research conducted based on the authors' descriptions of their analyses in Botswana only.	The article did not report on a method or application of R programming research conducted in Botswana
The research was added for review if more than one were in the same research area.	The article did not focus on Botswana.
The article was available in English.	The full text of the article was not available (for example, conference or abstracts).
The article was published between 2022 and 2005.	If research were commentaries and essays.

2.2. Data Extraction and Analysis Plan

For each article, data was extracted regarding: (i) the aim of research; (ii) research area; (iii) research conducted in Botswana; and (iv) R packages used. To analyse the data, a narrative review synthesis method was selected to capture the large range of R packages applied and whether the work was conducted in Botswana. It should be noted that a meta-analysis was not appropriate for this review given the broad range of research areas, R packages, and types of statistical models used in the research identified.

3. RESULTS

3.1. Overview of Article Characteristics

The search strategies using a combination of search terms identified 2300 articles that included a search term from the category in their abstract or title. The range for publication year of relevant articles was found to be between 2022-2018.A total of 10 articles were duplicates. Abstracts of 1090 articles were read by the author to perform an initial screening of eligibility for this scoping review. Of these, 620 were excluded because for not using R programming in their research work. 470 articles existed in one or more research area, however, 150 records were excluded for only existing in one research area which was not going to be efficient for reviewing from a sample research area. A total of 320 articles were selected for full-text review, but 302 records excluded for being more than 20 years except for one paper that had met all criteria which was [12]. This resulted in a total sample of 18 studies. The selected 18studies were reviewed in full by the author. In the subsequent narrative analysis, the paper focuses on the 18 studies that conducted research in Botswana that used R programming and R packages for their data analysis. Figure 1 below shows the PRISMA (Preferred Reporting Items for Systematic Reviews and MetaAnalyses) procedural flowchart on the process of identifying relevant research using R programming in Botswana.

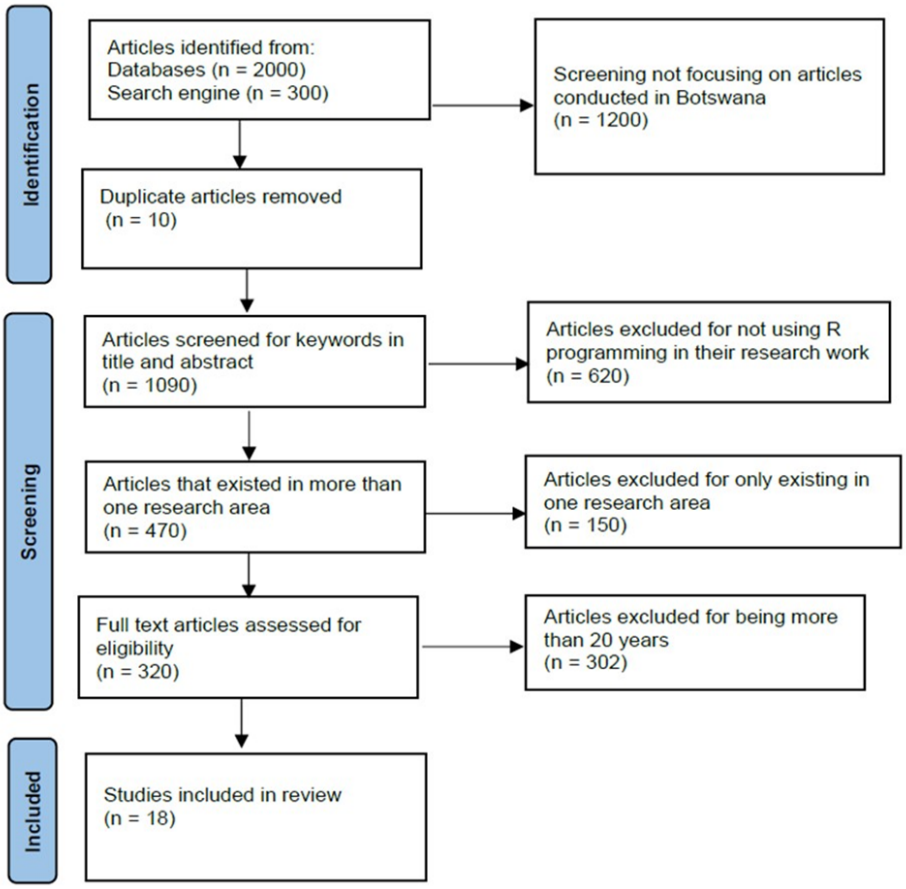

Figure 1. PRISMA (Preferred Reporting items for systematic Reviews and Meta- Analyses) procedural flowchart on R programming in Botswana

4. RESEARCH AREAS IN BOTSWANA THAT HAVE USED R PROGRAMMING

Through synthesis of the articles, four research areas conducted in Botswana using R programming and R programming applications were identified in figure 2 below.

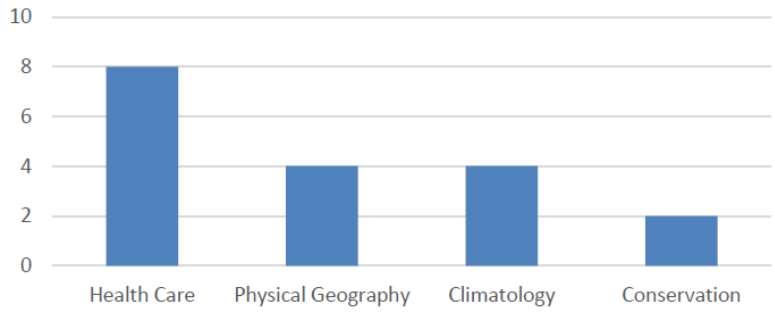

Figure 2. Research Areas in Botswana that have used R programming over the past 5 years

4.1. Health Care

Health Care included research that aim to identify, evaluate and improve health conditions in Botswana. There has been numerous research conducted in Botswana in order to improve the

efficiency of health facilities by addressing potential barriers. [14] has said Botswana is implementing an ambitious universal health coverage agenda and successfully expanding antiretroviral treatment to nearly 380,000 people living with human immunodeficiency virus (HIV). However, the country needs to critically assess its efficient use of all available resources to sustain gains and continue progress to attain the targets and vision for ending acquired immune deficiency syndrome (AIDS) as a public health threat by 2030. Descriptive analysis from data collected from regions of Botswana has helped to understand the actuality of treatment distribution, on-going treatment research and health institution management. For example, in an impact community breast examination study by[15] used records from a large clinical breast examination (CBE) initiative by a Botswana-based NGO. They sought to determine screening uptake, prevalence of breast abnormalities, number screened per breast cancer diagnosis, and clinical resources required to achieve diagnoses. Secondary analyses included proportion of women completing diagnostic evaluation and time to diagnosis. Their findings may inform planning of national CBE screening programming in Botswana and similar settings. Other research like [16] and[17] investigated exposure of HIV and TB of non-infected HIV people among HIVinfected people who have TB to determine infection rates within a community. COVID-19 research has also been conducted, however, there has only been one study by [18] as far as what has been identified that has used R programming for time series modelling of COVID-19 cases in Botswana in a 60-day period. The following section details the R packages used in Health care by the reviewed articles.

4.1.1. R Packages used in Health Care

4.1.1.1. Forecast Package

The Forecast package is part of the forecasting bundle by[19] which also contains the packages fma, expsmooth and Mcomp. The Forecast package contains functions for univariate forecasting, while the other three packages contain large collections of real time series data that are suitable for testing forecasting methods. The Forecast package implements automatic forecasting using exponential smoothing, ARIMA models, Theta method, cubic splines, as well as other common forecasting methods. [18] adopted a machine learning–based time series models, auto regressive integrated moving average (ARIMA) and Exponential smoothing algorithm or error trend season (ETS) to forecast confirmed COVID-19 cases in Botswana in a 60-day period. To successfully implement the ETS model, the Forecast package was executed in R.

4.1.1.2. Oce Package

[20]explained the Oce package simplifies Oceanographic analysis by handling the details of discipline-specific file formats, calculations, and plots. Designed for real-world application and developed with open-source protocols, Oce supports a broad range of practical work. Generic functions take care of general operations such as sub-setting and plotting data, while specialized functions address more specific tasks such as Hydrographic analysis, ADCP coordinate transformations, etc. It is easy to document work done with Oce, because its functions automatically update processing logs stored within its data objects. [21] presented an extended version of a general estimating equation-based approach for spatially correlated, binary data of HIV surveillance based on a pairwise composite likelihood that can accommodate penalized spline estimators. In addition, they applied it to antenatal HIV surveillance data collected in 2011 in Botswana to estimate the effects of proximity to the ‘hotspot’ of the country’s HIV epidemic and age on HIV prevalence. They obtained the global positioning system (GPS) coordinates of the administrative centre of each health district before applying Vincenty’s ellipsoidal formula for geodesic distance as implemented by the oce package.

4.1.1.3. Ggplot2 Package

According to [22], Ggplot2 is a framework that allows making both graphic plots and annotations. Ggplot2 enables developers to build almost any type of plot from dendrograms, network graphs and histograms. The R graph gallery explains that the R package is dedicated to data visualization and can realty improve the quality of the graphics.[23] sought to characterize genetic variation and to assess population substructure within a cohort of HIV-positive children from Botswana that is regionally underrepresented in genomic databases. They used annotations in their analysis that were visualized using Ggplot2.

4.1.1.4. SNP Relate Package

In a SNP Relate tutorial by [24], he described SNP Relate(R package formulti-core symmetric multiprocessing computer) is used in Genomic research due to the packages' two key computations on SNP data: principal component analysis (PCA) and relatedness analysis using identity-by-descent measures. SNP Relate provides a binary format for single-nucleotide polymorphism (SNP) data in Genome-wide association research (GWAS) utilizing Core Array Genomic Data Structure (GDS) data files. The GDS format offers the efficient operations specifically designed for integers with two bits, since a SNP could occupy only two bits. In [23], to assess substructure within the Botswana cohort, they followed the same QC pipeline as the PCA analysis. Independent autosomalmarkers pruned by LD (r-squared coefficient of 0.2) in windows of 1,000 base pairs advanced 100 SNPs at a time were used for the analysis with the SNP Relate package.

4.1.1.5. Inctools Package

[25]explained the Inctools package for estimating incidence from biomarker data in cross-sectional surveys, and for calibrating tests for recent infection. [26] further explained in their vignette of the package that it is broadly conceived to provide state of the art functionality to support numerous aspects of population level incidence surveillance. Inspiration for the work of the package derived from the challenges associated with estimating population level HIV incidence.[27] sought to determine HIV incidence in this setting with both high HIV prevalence and high ART coverage in Botswana. Statistical analysis and estimation of HIV incidence was implemented using the Inctools package.

4.1.1.6. Ape Package

In an analysis of Analyses of Phylogenetics and Evolution (APE) in R language by [28], the package is explained to be used in molecular evolution and phylogenetics. APE provides utility functions for reading and writing data and manipulating phylogenetic trees, as well as several advanced methods for phylogenetic and evolutionary analysis, for example, comparative and population genetic methods. APE takes advantage of the many R functions for statistics and graphics and also provides a flexible framework for developing and implementing further statistical methods for the analysis of evolutionary processes.[17] aimed to inform public health interventions by revealing the map of circulating HIV lineages (molecular HIV clusters) in Botswana and determining the extent of viral lineages spread within single and multiple communities. As part their method, multiple sequence alignment was generated by aligning near full-length HIV-1C sequences to the majority HIV-1Cconsensus sequence using mafft(multiple sequence alignment software) followed by removing positions with more than 98% gaps using the del.colgaps only() function from the APE package. Apart from using the APE package for their multiple segment methods, all confidence intervals of estimated proportions of asymptotic 95% binomial confidence intervals

(95% CI) were computed in Base R. Th associations between paired samples were tested by estimating Pearson correlation or Spearman rank correlation in Base R as well.

4.1.1.7. Adephylo Package

The Adephylo package is dedicated to the analysis of comparative evolutionary data. Phylogenetic comparative methods initially aimed at accounting for or removing the effects of phylogenetic signal in the analysis of biological traits. However, existing approaches have shown that considerable information can be gathered from the study of the phylogenetic signal. In particular, close examination of phylogenetic structures can unveil interesting evolutionary patterns. For this purpose, [29] developed the Adephylo package that provides tools for quantifying and describing the phylogenetic structures of biological traits. Adephylo implements tests of phylogenetic signal, phylogenetic distances and proximities, and novel methods for describing further univariate and multivariate phylogenetic structures. These tools open up new perspectives in the analysis of evolutionary comparative data. As part of the method to determine HIV viral lineage spread by [17], they used the distTips() function from the Adephylo package to calculate three pairwise distances per each molecular HIV-1 cluster identified by phylogenetic inference: mean Tamura Nei 93 (TN93)-corrected pairwise distances mean ML-corrected pairwise distances estimated by Randomized Accelerated Maximum Likelihood (RAxML) and median patristic distances inferred from the RAxML tree.

4.1.1.8. iGraph Package

[30] and [31] explained the iGraph package provides handy tools for researchers in network science. It is an open source portable library capable of handling huge graphs with millions of vertices and edges and it is also suitable for grid computing. It contains routines for creating, manipulating and visualizing networks, calculating various structural properties, importing from and exporting to various file formats and many more. Via its interfaces to high-level languages like GNU (GNU's Not Unix! software) R and Python, it supports rapid development and fast prototyping. For [17] to analyse the spread of identified phylogenetically distinct HIV-1C lineages across communities by enumerating links (edges)within molecular clusters, viral links within molecular HIV clusters were visualized on a map of Botswana using the iGraph package.

4.1.1.9. Base R

Base R contain a set of standard (Base) packages which are considered part of the R source code and automatically available as part of R installation. Base packages contain the basic functions that allow R to work and enable standard statistical and graphical functions on datasets. Not only did the reviewed articles use R packages, but they also used Base R to carry out their statistical models. Under Health Care, [15] evaluated a clinical breast examination (CBE) screening program to determine the prevalence of breast abnormalities, number examined per cancer diagnosis, and clinical resources required for these diagnoses in rural and urban Botswana. All tests were two-tailed with a significance level of 0.05. Analysis was performed using Base R. [16] compared exposure to TB patients between HIV-infected and non-HIV infected health care workers (HCW) in health facilities in Botswana. Their statistical analysis was performed using Base R. No further details were revealed as to which functions were used for a specific operation. [18] used Akaike information criterion (AIC) assessment metric to select an appropriate model automatically using the principle of maximum likelihood estimation. In this phase, the simplest ETS model, ETS (A, Ad, N), described by an Additive error, Additive trends and No seasonal patterns, is chosen among 30

possible state space models. To successfully execute AIC assessment metric, a grid search algorithm (GSA) was implemented using Base R as well.

4.2. Physical Geography

Physical Geography comprised of research in Geomorphology, Hydrology, and also Biogeography as practiced by geographers, such as [13] that aimed to measure the change progression of the physical environment and measure the distribution of organisms. Physical geography makes up natural resources and wildlife, and Botswana has generated a lot of research for sustainability such as the Okavango Research Institute (ORI) of the University of Botswana on wetlands and adjacent drylands. The institute undertakes engaged research and provides training and service in wetland and adjacent dryland social and ecological systems with support from the Southern African Science Service Centre for Climate Change Adaptive Land Management (SASSCAL).

As it is written under the section 4, Physical Geography includes Geomorphology, Hydrology, Ecology and also Biogeography. The following section details the R packages used in Physical Geography by the reviewed articles.

4.2.1. R Packages used in Physical Geography

4.2.1.1. Metab Rpackage

Metab R is for high-throughput analysis of metabolomics data generated by GC–MS. MetabR was developed by the R Development Core Team in 2008. The package uses the R programming language for data correction, filtering and reshaping of datasets initially produced by AMDIS[32]. [32] combined the power of AMDIS, Metab R and MINITAB to classify the floral geographical origins of three randomly selected commercially produced and three unprocessed natural organic honeys from Zambia and Botswana using GC–MS untargeted metabolomics of volatile components. [32] used the Metab R package to recalculate peak intensities, display peak areas, remove false positives, normalize by internal standard, normalize by biomass as well as carry out the h test, that is, analysis of variance (ANOVA) and the t test on data generated by AMDIS for metabolomics, both targeted and non-targeted.

4.2.1.2. NmlePackage

The Nmle package was developed for fitting and comparing gaussian linear and nonlinear mixed-effects models[33]. The online tutorial on using R for reproducible research by [34] explained that the packagelets you specify variance-covariance structures for residuals and is well suited for repeated measure or longitudinal designs. They further explain that it contains sample data, statistical functions, matrices, and a lattice framework.[35] classified multi-decadal changes in land cover in the semi-arid Chobe District in northeastern Botswana using post-classification analysis of Landsat datasets from 1990, 2003, and 2013.A two-sample t-test of equal variances was used to test the hypothesis that mean fire occurrence was significantly different in alternating (even and odd) years, and ordinary least squares (OLS) regression was used to analyze the association between total annual rainfall and fire count. Here, they fit generalized least squares (GLS) regressions to these data in the Nmle package using five autocorrelation structures: exponential, Gaussian, spherical, linear, and rational quadratic.

4.2.1.3. Caret package

The Caret package, short for classification and regression training, contains numerous tools for developing predictive models using the rich set of models available in R. The package focuses on simplifying model training and tuning across a wide variety of modeling techniques. It also includes methods for pre-processing training data, calculating variable importance, and model visualizations. [35] calculated overall classification accuracy with binomial 95% confidence intervals for net LCC using the Caret package by dividing the total number of correctly identified reference points by the total number of sample units in the matrix.

4.2.1.4. Rpart package

According to [36], the Rpart package functions build classification or regression models of a very general structure using a two stage procedure and the resulting models can be represented as binary trees. An example is some preliminary data gathered at Stanford on revival of cardiac arrest patients by paramedics. The goal was to predict which patients can be successfully revived in the field based on fourteen variables known at or near the time of paramedic arrival, e.g., sex, age, time from attack to first care, etc. Since some patients who are not revived on site are later successfully resuscitated at the hospital, early identification of these "recalcitrant" cases is of considerable clinical interest. [37] pooled training methods for classifying open water, water with emergent vegetation, and non-flooded land cover in the Chobe River Basin study area from selected moderate resolution imaging spectroradiometer (MODIS) images drawn from each of three years from 2014 to 2016. All calculations and classifications were done using the Rpart package.

4.2.1.5. Base R

Base R was also used in some Physical geography research such as [38] who investigated the relationship among woody vegetation, precipitation, grazing, borehole density, and fire. All regression was analysed using Base R to explore the environmental drivers of woody species richness and species abundances. All statistical analysis in [35]was conducted using Base R programming.

4.3. Conservation

Conservation included research that aimed to evaluate preservation practices for wildlife in Botswana. [38] wrote that the transition of savanna ecosystems to open shrubland across Botswana and in particular the western part of the Kalahari, presents a considerable threat to the conservation of the economically important ranching industry. In order to develop adaptive management strategies, the underlying environmental drivers of woody vegetation species need to be better understood. By understanding the environmental drivers responsible for the diversity and abundance of woody vegetation, develop predictive models need to be developed to identify 'high-risk' areas, and provide managers, farmers, and governments with decision support across savanna landscapes. Previous research addressing the ecological processes responsible for the observed vegetation patterns have often found conflicting results regarding the importance and significance of these environmental drivers. The following section details the R packages used in Conservation.

4.3.1. R Packages used in Conservation

4.3.1.1. Cluster Package

The Cluster package was developed by [39] which contain cluster methods for Cluster analysis to find groups in data.[40] evaluated banded mongoose den site use and attributes across anthropogenic and natural landscapes in Northern Botswana and discuss implications for disease transmission in changing landscapes. For their cluster analysis, the Cluster package and specifically the Daisy algorithm, were used for hierarchical cluster analysis and creation of cluster dendrograms of den site variables. Using the sil_width function, the entire clustering was displayed by combining the silhouettes, which represents clusters, into a single plot and the average silhouette width provided a means to select the appropriate number of clusters.

4.3.1.2. Rafalib Package

The Rafalib package contains a series of shortcuts for routine tasks originally developed to facilitate data exploration [41].[40] used the Rafalib package and the myplclust function to plot the remaining dendrogram from a set of dendrograms.

4.3.1.3. RpartPackage

Considering the description of the Rpart package in section 4.2.1.4, the Rpart package was also used in Conservation research by [40] when they used classification and regression (CART), a nonparametric decision tree method to evaluate associations between land use and banded mongoose den sites. They used the Rpart package and the class method to create the decision trees. Land use was set as the dependent variable and selected the lowest error (estimate of the cross-validation prediction error) and the corresponding conditional probability (cp) in the Rpart object to identify the optimal level for pruning the resulting tree.

4.3.1.4. PartykitPackage

[42] said the Partykit package provides a flexible toolkit for learning, representing, summarizing, and visualizing a wide range of tree-structured regression and classification models. The functionality encompasses: (a) basic infrastructure for representing trees (inferred by any algorithm) so that unified print/plot/predict methods are available. (b) dedicated methods for trees with constant fits in the leaves (or terminal nodes) along with suitable coercion functions to create such trees (e.g., by Rpart, RWeka, PMML). And (c), a reimplementation of conditional inference trees (ctree, originally provided in the party package); (d) an extended reimplementation of model-based recursive partitioning (mob, also originally in party) along with dedicated methods for trees with parametric models in the leaves. The Partykit provides a common unified infrastructure for recursive partitioning in the R system for statistical computing. In particular, Partykit package provides tools for representing, printing, plotting trees and computing predictions.[40] used the as.party function from in the Partykit package to plot their final decision tree.

4.3.1.5. BaseR

[43] received help from David R Roberts for help coding in R in theory study on comparing two widely practiced standards for counting animals - aerial strip surveys and ground line transects - with interpreted counts of animal tracks in the Kalahari. However, the study does not describe how R programming was used for and which R packages the study uses.[40]

used the fisher test function in Base R to perform the Fisher's exact test to determine differences between anthropogenic and natural den structures and their use during study periods conducted. All their other statistical analysis was conducted in Base R as well.

4.4. Climatology

Climatology research included measuring the change in rainfall and rainfall composition over a period of time. Climate change in Botswana can be contributed by the Okavango Delta. The Okavango Delta consists of permanent and seasonal wetlands bordering onto dry, occasionally flooded grassland and forest areas.

4.4.1. R Programming Packages In Climatology

All the reviewed research identified to have used Climatology has not used R packages for their work.[44] computed R/S analysis for the residual series obtained after filtering periodicities,showed that they have statistical properties close to the ordinary Brownian noise using Base R.A study by [45] used fluxes calculated using Base R to assess net atmospheric Methane (CH_4) fluxes and controlling environmental factors in the Okavango Delta.[46] analysed rainfall concentration and spatio-temporal trends in annual and seasonal November to March rainfall in Botswana. The Precipitation Concentration Index (PCI), Mann–Kendall trend test, Theil–Sen's slope estimator (β), Autocorrelation Function (ACF) and relative percentage change (RPC) methods were adopted for data analysis were performed using Base R. [47] research determined trends in the annual maximum average rainfall for Botswana during the years 1901-2012. All their descriptive statistics, trend analysis and change point detection techniques were developed using Base R.

5. DISCUSSION

This paper aimed to synthesis Botswana research that has applied R programming for their data analysis. Research done in Botswana that have used R programming were identified in four research areas Health Care, Physical Geography, Conservation and Climatology. Predominantly, research in Botswana that have used R programming have used it for statistical, descriptive analysis and exploratory analysis to demonstrate the patterns in weather, national program evaluations and change in the landscape. The variety of data visualization, mapping, bio markering and genomic operations used explain the many R packages used in the health care research. The reason for R packages used in certain research areas may be due to certain R packages being made for particular operations, for example, SNP Relate and Ape packages are for Genomics and ADE package is created for Phylogenetics for health care. The most used R package across the different research areas is the Rpart package which may be due to its methods and operations for classification modeling as shown in figure 3. Classification showed to be the most used technique in the reviewed research with Nmle, Caret, Rpart, Forecast and Partykit R packages used for classification, regression and gaussian modeling as shown in figure 4 below. This proves that R programming can be used for machine learning contrary to the belief that machine learning is more appropriate for Python programming.

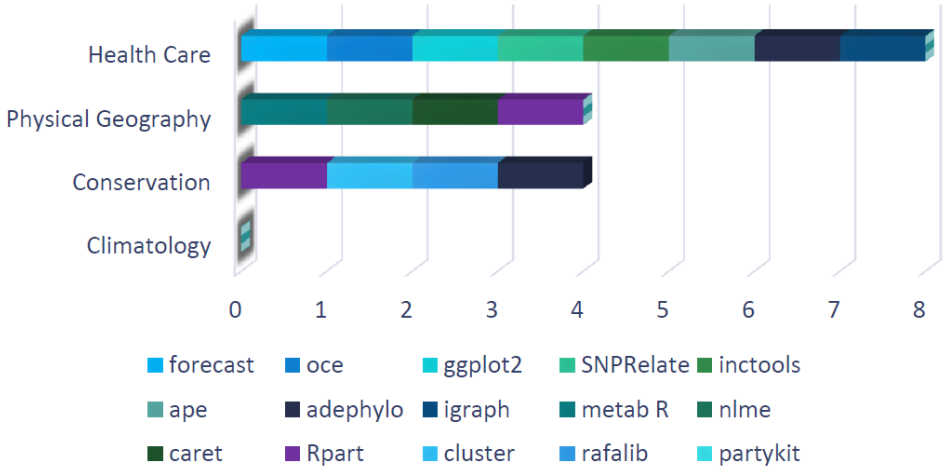

Figure 3. R Packages used per Research Area

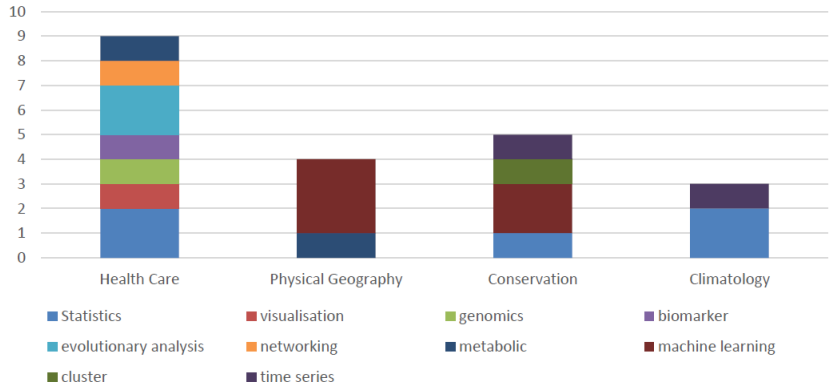

Figure 4. Techniques used per Research Area

As R is an open-source software, it is open to developers to create packages that have operations to perform a certain functions. Recently, a team from the University of Mannheim developed an R package called, Overview R for exploratory data analysis that gives an overview of data and identifies where the gaps are in data with a particular emphasis on a time series cross-sectional consideration[48]. Another R package called STEM (Spatio-temporal models) was developed by Professor Michela Cameletti from the department of Economics at the University of Bergamo, Italy. The Stem package uses estimation of the parameters of a spatio-temporal model using the EM algorithm, estimation of the parameter standard errors using a spatio-temporal parametric bootstrap and spatial mapping. The creation and updating of R packages are encouraged and are updated in The Comprehensive R Archive Network (CRAN).All the R packages created that have been used in research have been by developers outside of Africa. To the best of the authors' knowledge, there are currently no documented R packages created by R practitioners from Botswana or Africa. However, the R packages used are necessary for carrying out a particular operation such as classification and regression using Metab R. Research in Climatology did not use R packages but instead, they used Base R functions that are fundamental R functions or methods for data analysis, manipulation, data cleaning, and data formatting.

5.1. Limitations

This review paper focused on the research conducted in Botswana that used R programming for data analysis and did not compare other research across Africa that use R programming. To the best of the authors' knowledge, there is a gap in Botswana research that use R programming in Botswana from different research areas which maybe due to the limited R experts who are proficient in R programming. [43] had assistance for coding in R statistical environment and for assistance with graphics from programmers from other universities outside Botswana. A reason to the limited R proficiency may be that in academia, SPSS, STATA and ATLAS-TI are commonly taught for data science, they are more accepted and are still popularly used for descriptive statistical analysis. In Botswana, very few people know about R programming for data science and for project work and those who do know it, still consider it a statistical computing software even though it can be used for predictive modeling with the TidyModel and Caret packages. This review paper also did not look at the statistical models used by each study due to the number of models per study. This can be used as future work to consider which statistical models are appropriate for a research area.

6. CONCLUSIONS

In summary, this paper set out to use a scoping review method to synthesis research conducted in Botswana that use R programming in their data analysis. R programming is an open-source software which means developers can share and add their developed R packages to the R Development Core Team for distribution. The paper followed the PRISMA methodology for a systematic review. It was found that the few research conducted in Botswana that use R programming may be due to the few expert R programming practitioners and the research areas found were in Health Care, Climatology, Physical Geography and Conservation. A variety of R packages were mostly used in Health Care ranging from plotting, genomics and networking. Classification modelling was the common analysis used in the research which means R programming can be used for machine learning experiments contrary to the common belief. In addition, the growth of R in data science could be seen to be used in future machine learning research such as trading volume as a predictor of market movement using Logistic regression in the R environment.The statistical models were not considered in this review as well, however, a review of the use of the statistical analysis for decision making can be considered for future work.

REFERENCES

[1] Kaya, Efdal, Muge Agca, FatihAdiguzel, and Mehmet Cetin. "Spatial data analysis with R programming for environment." Human and ecological risk assessment: An International Journal 25,no. 6. 2019: pp. 1521-1530.

[2] Çetinkaya-Rundel, Mine, and Colin Rundel. "Infrastructure and tools for teaching computing throughout the statistical curriculum." The American Statistician 72, no. 1.2018: pp. 58-65.

[3] Sloane, Lori. "Library/Software/Data Carpentries." (2022).

[4] Government of Botswana. "Monuments and Relics, Act." 2001.

[5] Ama, Njoku O., and Charles M. Fombad. "Patent and research exemption: Challenges for research capacity and utilization in universities, research institutions and industry in Botswana." International Journal of Asian Social Science 1, no. 5. 2011: pp. 157-180.

[6] Republic of Botswana. "Botswana National Research, Science and Technology Plan Final report". Ministry of Communications, Science and Technology Gaborone, Botswana (2005).

[7] Baliyan, Som Pal, and Fazlur Rehman Moorad. "Teaching Effectiveness in Private Higher Education Institutions in Botswana: Analysis of Students' Perceptions." International Journal of Higher Education 7, no. 3. 2018: pp. 143-155.

[8] Matenge, Tjedza G., and Bob Mash. "Barriers to accessing cervical cancer screening among HIV positive women in Kgatleng district, Botswana: a qualitative study." PLoS One 13, no. 10 2018: e0205425.

[9] Ntshebe O, Channon AA, Hosegood V. "Household composition and child health in Botswana" BMC public health. 2019;19(1): pp. 1-3.

[10] Shatte, Adrian BR, Delyse M. Hutchinson, and Samantha J. Teague. "Machine learning in Mental health: a scoping review of methods and applications." Psychological medicine 49, no.9. 2019: pp.1426-1448.

[11] Saqib K, Khan AF, Butt ZA. Machine learning methods for predicting postpartum depression: Scoping review. JMIR mental health. 2021 Nov 24;8(11):e29838.

[12] Zhou, X., N. Persaud, and H. Wang. "Periodicities and scaling parameters of daily rainfall over semi-arid Botswana." Ecological modelling 182, no. 3-4.2005: pp. 371-378.

[13] Harden, C. P., Luzzadder-Beach, S., MacDonald, G. M., Marston, R. A., & Winkler, J. A."Physical geography contributes". Progress in Physical Geography: Earth and Environment, 44(1), 2020. pp.5-13.

[14] Avila, C., Zeng, W., & Cintron, C. "Efficiency of health facilities providing antiretroviral treatment services in Botswana". Journal of Hospital Management and Health Policy. 2020.

[15] Dykstra M, Malone B, Lekuntwane O, Efstathiou J, Letsatsi V, Elmore S, Castro C, Tapela N, Dryden-Peterson S. "Impact of community-based clinical breast examinations in Botswana".JCO Global Oncology. 2021;7: pp. 17-26.

[16] Shin, Sanghyuk S., ChawangwaModongo, Nicola M. Zetola, Qiao Wang, Thabo Phologolo, Mary Kestler, and Ari Ho-Foster. "High rates of exposure to tuberculosis patients among HIV-infected health care workers in Botswana." The international journal of tuberculosis and Lung disease 22, no. 4. 2018: pp. 366-370.

[17] Novitsky V, Zahralban-Steele M, Moyo S, Nkhisang T, Maruapula D, McLane MF, LeidnerJ,Bennett K, Wirth KE, Gaolathe T. "Mapping of HIV-1C transmission networks reveals extensive spread of viral lineages across villages in Botswana treatment-as-prevention trial."The Journal of infectious diseases. 2020 ;222(10): pp. 1670-80.

[18] Mphale, Ofaletse, and V. Lakshmi Narasimhan. "Comparative Forecasts of Confirmed COVID-19 Cases in Botswana Using Box-Jenkin's ARIMA and Exponential Smoothing State-Space Models." In Recurrent Neural Networks, pp. 355-381.2022. CRC Press.

[19] Hyndman, R. J., &Khandakar, Y. "Automatic time series forecasting: the forecast package for R." Journal of statistical software, 27, 2008. pp. 1-22.

[20] Kelley, Dan E. "The OCE package." In Oceanographic Analysis with R, pp. 91-101. Springer,New York, NY, 2018.

[21] Wirth, Kathleen E., et al. "A composite likelihood approach for estimating HIV prevalence in the presence of spatial variation." Statistics in medicine 34.28 (2015): pp.3750-3759.

[22] Valero-Mora PM. "ggplot2: elegant graphics for data analysis." Journal of Statistical Software.2010;35: pp.1-3.

[23] Retshabile G, Mlotshwa BC, Williams L, Mwesigwa S, Mboowa G, Huang Z, RustagiN,Swaminathan S, Katagirya E, Kyobe S, Wayengera M. "Whole-exome sequencing reveals uncaptured variation and distinct ancestry in the Southern African population of Botswana." The American Journal of Human Genetics. 2018 ;102(5):pp.731-743.

[24] Zheng X. "A tutorial for the R Package SNPRelate." University of Washington, Washington,USA.2013.

[25] Welte, A., E. Grebe, A. McIntosh, P. Bäumler, R. Kassanjee, and H. Brand. "inctools: incidence estimation tools." 2017.

[26] Eaton J, Grebe E, Baumler P, McIntosh A, Ongarello S, Welte A, Kassangee R, Brand H, Van Schalkwyk C, Li Y, Daniel S. "Incidence Estimation Tools (inctools)."2017.

[27] Moyo S, Gaseitsiwe S, Mohammed T, Pretorius Holme M, Wang R, Kotokwe KP, Boleo C, Mupfumi L, Yankinda EK, Chakalisa U, Van Widenfelt E. "Cross-sectional estimates revealed high HIV incidence in Botswana rural communities in the era of successful ART scale-up in 2013-2015." PLoS One. 2018 ;13 (10):e0204840.

[28] Paradis E, Claude J, Strimmer K. "APE: analyses of phylogenetics and evolution in R language.Bioinformatics." 2004 Jan 22;20(2):289-90.

[29] Jombart T, Dray S. "Adephylo: exploratory analyses for the phylogenetic comparative method.Bioinformatics." 2010 Apr 5;26(15):1-21.
[30] Csardi G, Nepusz T. "The igraph software package for complex network research." InterJournal,complex systems. 2006 ;1695(5):1-9.
[31] Csardi MG. Package 'igraph'. Last accessed. 2013 Apr 7;3(09):2013.
[32] Sichilongo K, Padiso T, Turner Q. "AMDIS-Metab R data manipulation for the geographical and floral differentiation of selected honeys from Zambia and Botswana based on volatile chemical compositions using SPME–GC–MS." European Food Research and Technology. 2020;246(8):1679-90.
[33] Pinheiro J. "nlme: linear and nonlinear mixed effects models." R package version 3.1-96. 2009 [Online]. Available: http://cran. r-project. org/web/packages/nlme/.
[34] Abdelrahman, K., Contreras, A., Degtyarev, Z., Deng, J., Foster, J., Franz, A., Funderburk, T., Horger, M., Kravitz, J., Lakshin A., Manigat, M., Trois, R., Vasquez, A., Vo, A., Wilson, N., and Yeremenko, M. "Using R for Reproducible Research": Student Contributed Tutorials. (M. J. C. Crump, Ed.).2019. [Online]. Available: https://crumplab.com/psy7709_2019/book/docs.
[35] Fox JT, Vandewalle ME, Alexander KA. "Land cover change in northern botswana: the influence of climate, fire, and elephants on semi-arid savanna woodlands." Land. 2017 ;6(4):73.[36]
[36] Therneau TM, Atkinson EJ. "An introduction to recursive partitioning using the RPART routines." Mayo Foundation: Technical report; 1997.
[37] Braget MP, Goodin DG, Wang J, Hutchinson JM, Alexander K. "Flooded area classification Using pooled training samples: an example from the Chobe River Basin, Botswana." Journal of Applied Remote Sensing. 2018 ;12(2):026033.
[38] Meyer T, Holloway P, Christiansen TB, Miller JA, D'Odorico P, Okin GS. "An assessment of multiple drivers determining woody species composition and structure: A case study from the Kalahari, Botswana." Land. 2019 ;8(8):122.
[39] Kaufman L, Rousseeuw PJ. "Finding groups in data: An introduction to cluster analysis"–john wiley& sons. Inc., New York. 1990.
[40] Nichols CA, Alexander KA. "Characteristics of banded mongoose (Mungos mungo) den sites Across the human-wildlife interface in Northern Botswana." Mammalian Biology. 2019 ;97(1):80-7.
[41] Irizarry, R. A., and M. I. Love. "rafalib: Convenience Functions for Routine Data Exploration.R package version 1.0. 0." (2015).
[42] Hothorn T, Zeileis A. "partykit: A modular toolkit for recursive partytioning in R." The Journal Of Machine Learning Research. 2015 ;16(1):3905-9.
[43] Keeping D, Burger JH, Keitsile AO, Gielen MC, Mudongo E, Wallgren M, Skarpe C, Foote AL. "Can trackers count free-ranging wildlife as effectively and efficiently as conventional aerial survey and distance sampling? Implications for citizen science in the Kalahari, Botswana." Biological Conservation. 2018; 223:pp.156-169.
[44] Zhou X, Persaud N, Wang H. Periodicities and scaling parameters of daily rainfall over semi-Arid Botswana. Ecological modelling. 2005 ;182(3-4):371-8.
[45] Gondwe MJ, Helfter C, Murray-Hudson M, Levy PE, Mosimanyana E, Makati A, MfundisiKB,Skiba UM. Methane flux measurements along a floodplain soil moisture gradient in the Okavango Delta, Botswana. Philosophical Transactions of the Royal Society A. 2021 ;379(2210):20200448.
[46] Gökçekuş H, Kassem Y, Mphinyane LP. "Analysis of Spatio-temporal rainfall trends and Rainfall variability in Botswana between 1958 and 2019." International Advanced Researches and Engineering Journal. 2021;5(3): pp.444-453.
[47] Thupeng WM, Thekiso TB. "Changepoint analysis: A practical tool for detecting abrupt changes in rainfall and identifying periods of historical droughts: A case study of Botswana." Bull. Math. Stat.Res. 2019;7: pp. 33-46.
[48] Meyer C, Hammerschmidt D. overviewR-Easily Explore Your Data in R. Journal of Open Source Software. 2022 ;7(77):4740.

AUTHOR

Simisani Ndaba is a Teaching Assistant in the Department of Computer Science at the University of Botswana. She hasa Masters of Science in Computer Information Systems where her research work was based in Information Retrieval. She also holds a Bachelor's degree in Business Information Systems and a Post Graduate Diploma in Education in Computer Science.