Assessment 1: Research Proposal

Heuristic Approach to Handling Anomalies in Data

|  |  |
| --- | --- |
| Name: | Supervisor: |
| University ID: |  |
| Word Count: 2687 (Without References) |  |
| Date |  |

Table of Contents

[Table of Figures 2](#_Toc69065782)

[Abstract 3](#_Toc69065783)

[Introduction 3](#_Toc69065784)

[Background 4](#_Toc69065785)

[Aim/Objective 4](#_Toc69065786)

[Research Questions 4](#_Toc69065787)

[Ethical Considerations 4](#_Toc69065788)

[Literature Review 4](#_Toc69065789)

[Survey on Anomaly Detection using Data Mining Techniques 5](#_Toc69065790)

[Unsupervised real-time anomaly detection for streaming data 6](#_Toc69065791)

[A Comparative Evaluation of Unsupervised Anomaly Detection Algorithms for Multivariate Data 10](#_Toc69065792)

[Project Timeline 11](#_Toc69065793)

[Methodology 12](#_Toc69065794)

[Project Evaluation 12](#_Toc69065795)

[Conclusion 12](#_Toc69065796)

[References 12](#_Toc69065797)

# Table of Figures

[Figure 1 - Methodology for Anomaly Detection 5](#_Toc69065798)

[Figure 2 - Papers with Hybrid Approaches to Anomaly Detection 6](#_Toc69065799)

[Figure 3 - Real world Temperature Sensor Data with Anomalies marked as Dark circles. 7](#_Toc69065800)

[Figure 4 - Methodology to Create an Anomaly Detector based on HTM 7](#_Toc69065801)

[Figure 5 - HTM (Hierarchical Temporal Memory Workflow) 8](#_Toc69065802)

[Figure 6 - Results of Different Algorithms with NAB Dataset 9](#_Toc69065803)

[Figure 7 - Properties Comparison for Different Algorithms on the NAB Dataset 9](#_Toc69065804)

[Figure 8 - Real-time Detection Results of Different Algorithms on the NAB Dataset 9](#_Toc69065805)

[Figure 9 - Different Anomaly Detection Methodologies Based on the Machine Learning Style 10](#_Toc69065806)

[Figure 10 - Datasets used for comparison in this Research 10](#_Toc69065807)

[Figure 11 - Results for All the Algorithms for the Datasets used. 11](#_Toc69065808)

[Figure 12 - Project Timeline 11](#_Toc69065809)

# Abstract

The world is seeing a gigantic expansion in the accessibility of streaming, time-arrangement information. Generally determined by the ascent of associated ongoing information sources, this information presents specialized difficulties and openings. One principal ability for streaming examination is to demonstrate each stream in an unaided style also, recognize strange, abnormal practices continuously. Early irregularity identification is important, yet it tends to be hard to execute dependably practically speaking. Application limitations expect frameworks to deal with information in real-time, not groups. Streaming information naturally displays idea float, preferring calculations that adapt continuously. Moreover, the huge number of free streams practically speaking necessitates that peculiarity finders be completely computerized.

Anomaly detection is the way toward recognizing unforeseen things or occasions in datasets, which contrast from the standard. Rather than standard order assignments, abnormality discovery is frequently applied on unlabelled information, taking just the inward design of the dataset into account. This test is known as unaided oddity discovery and is tended to in numerous viable applications, for instance in network interruption discovery, misrepresentation identification as well as in the existence science and clinical area. Many calculations have been proposed in this territory, yet shockingly the examination local area actually does not have a relative widespread evaluation just as regular freely accessible data.

The anomaly this research will be focused on is the anomaly of missing values. This research will employ regression algorithms to predict the missing values in a dataset based on training from available data. The regression algorithm used will be ARIMA.

# Introduction

The location of irregularities progressively streaming information has viable and huge applications across numerous ventures. Use cases like safeguard support, extortion counteraction, deficiency recognition, and observing can be found all through various enterprises like money, IT, security, clinical, energy, online business, horticulture, and web-based media. Recognizing peculiarities can give significant data in basic situations, yet solid arrangements don't yet exist (Ahmad, Lavin, Purdy and Agha, 2017).

In AI, the identification of "not-ordinary" occurrences inside datasets has consistently been of incredible interest. This interaction is ordinarily known as oddity recognition or exception detection. “An outlying observation, or outlier, is one that appears to deviate markedly from other members of the sample in which it occurs” (Goldstein and Uchida, 2016). Albeit this definition is as yet legitimate today, the inspiration for recognizing these exceptions is totally different at this point.

In those days, the primary justification the discovery was to eliminate the exceptions a while later from the preparation information since design acknowledgment calculations were very touchy to exceptions in the information. This technique is likewise called information purging. After the advancement of something else powerful classifiers, the interest in inconsistency identification diminished a great deal. In any case, there was a defining moment around the year 2000, when specialists began to get keener on the oddities itself, since they are frequently connected with specific intriguing occasions or dubious information records (Goldstein and Uchida, 2016).

Today anomalies are known to have two significant qualities (Goldstein and Uchida, 2016):

1. Irregularities are not quite the same as the standard regarding their highlights.
2. They are uncommon in a dataset contrasted with typical occasions.

Abnormality identification calculations are presently utilized in numerous application spaces and regularly improve conventional guideline based recognition frameworks (Goldstein and Uchida, 2016).

## Background

The rate of development of technology in the IT sector has brought about an intense amount of data being generated. This data, although very concentrated both qualitatively and quantitatively, is becoming a lot to handle properly. This resulted in the advent of the automation of data collection. Albeit it is easy to leave the data collection to the automation process, the integrity and quality of the data must be ensured. There are many issues that occur when data collection is automated. These issues are called anomalies.

The most pressing issue is the missing data instances anomaly. Missing data harms the accuracy of the data analysis and Machine Learning significantly. To properly handle missing data, researchers are researching different and innovative ways to tackle and solve this issue.

## Aim/Objective

The aim of this research project is to restore missing data in a dataset using Regression algorithms. To this extent, these steps are followed:

* Literature Review of existing research done by peers in the same domain.
* Dataset selection and finalization.
* Data Pre-processing.
* Machine Learning Algorithm design.
* Performance Metrics using cross-validation and metrics for Regression.

## Research Questions

The following question gives reason to the project for moving forward:

1. How effective is Regression if used for solving the anomaly of missing data instances?

## Ethical Considerations

Ethics is a complicated subject that has only become more prominent during the advent of Big Data. The UK Data Service department also provides guidelines for ethical research with specific relation to Big Data (UK Data Service, 2020). These guidelines will form the basis for this reports ethical approach. Some of the concerns that will be addressed are:

* Maintaining confidentiality in line with Birmingham City University (BCU) and DC guidelines,
* Anonymising information that violates group privacy,
* Ensuring transparency in reasons for data collection,
* Ensuring data is only used for the direct purpose it has been requested,
* Referencing sources for all information used within the research project,
* Ensuring all data is stored in the correct location. DC information must remain on DC servers.

As this project encounters any further ethical concerns these will be met within the recommended UK guidelines and with the advice of BCU and DC supervising members.

## Literature Review

A strong literature review arises from reviewing the work done by fellow researchers in the same domain. This LR in-turn validates the research and provides integrity to the work put in. Following are the papers finalized for LR in this research.

### Survey on Anomaly Detection using Data Mining Techniques

This paper surveys different information mining strategies for irregularity recognition to give better comprehension among the current procedures that may assist intrigued specialists with working future toward this path.

Anomaly Detection is the way toward discovering the examples in a dataset whose conduct isn't typical on anticipated. These sudden practices are additionally named as peculiarities or exceptions. The peculiarities can't generally be arranged as an assault yet it can be an astounding conduct which is beforehand not known. It might be destructive. At the point when information needs to be investigated to discover relationship or to anticipate known or obscure information mining procedures are utilized.

These incorporate grouping, arrangement and machine-based learning strategies. Mixture approaches are additionally being made to achieve more significant level of exactness on recognizing oddities. In this methodology the creators attempt to consolidate existing information mining calculations to infer better outcomes. Along these lines recognizing the strange or surprising conduct or peculiarities will respect contemplate and arrange it into new sort of assaults or a specific kind of interruptions. This overview endeavours to give a superior comprehension among the different kinds of information mining approaches towards oddity location that has been made up to this point.

#### Methodology

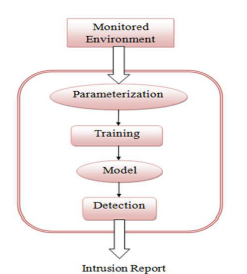


Figure - Methodology for Anomaly Detection

Abnormalities are patterns in the information that don't adjust to an all-around characterized ordinary conduct. The reason for inconsistency might be a noxious action or some sort of interruption. This unusual conduct found in the dataset is fascinating to the expert and this is the main component for Anomaly recognition.

In this paper different information digging procedures are depicted for the irregularity identification that had been proposed in the previous few years. This audit will be useful to scientists for acquiring an essential understanding of different methodologies for the irregularity discovery. Despite the fact that much work had been finished utilizing free calculations, cross breed approaches are by and large endlessly utilized as they give better outcomes furthermore, defeat the downside of one methodology over the other.

Consistently new obscure assaults are seen and accordingly there is a need of those methodologies that can identify the obscure conduct in the informational index put away, moved or adjusted. In this exploration work combination or mix of previously existing calculations are referenced that have been proposed. Intrigued specialists can consolidate the altered form of previously existing calculations. For instance, there are different new methodologies in the alteration of choice trees (like ID3, C4.5), GA, SVM (counting upgraded and numerous portion-based methodologies). This may yield more precise outcomes.

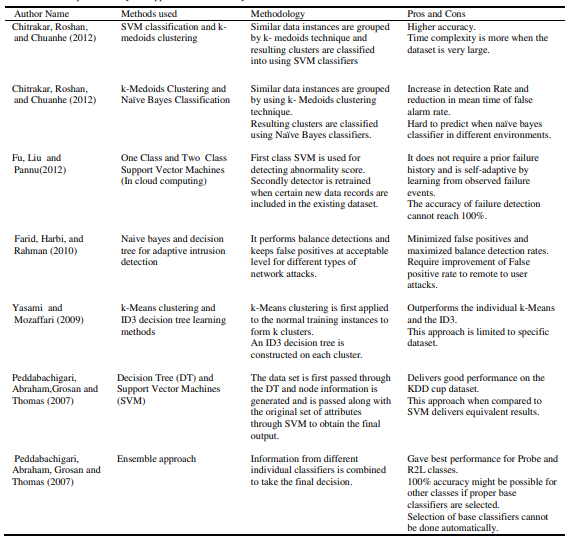


Figure - Papers with Hybrid Approaches to Anomaly Detection

### Unsupervised real-time anomaly detection for streaming data

In this paper, the authors propose a novel abnormality identification calculation that meets these imperatives. The method depends on an online grouping memory calculation called Hierarchical Temporal Memory (HTM). We likewise present outcomes utilizing the Numenta Anomaly Benchmark (NAB), a benchmark containing certifiable information streams with marked abnormalities. The benchmark, the first of its kind, gives a controlled open-source climate for testing oddity recognition calculations on streaming information. The authors present outcomes and investigation for a wide scope of calculations on this benchmark, and talk about future difficulties for the arising field of streaming examination.

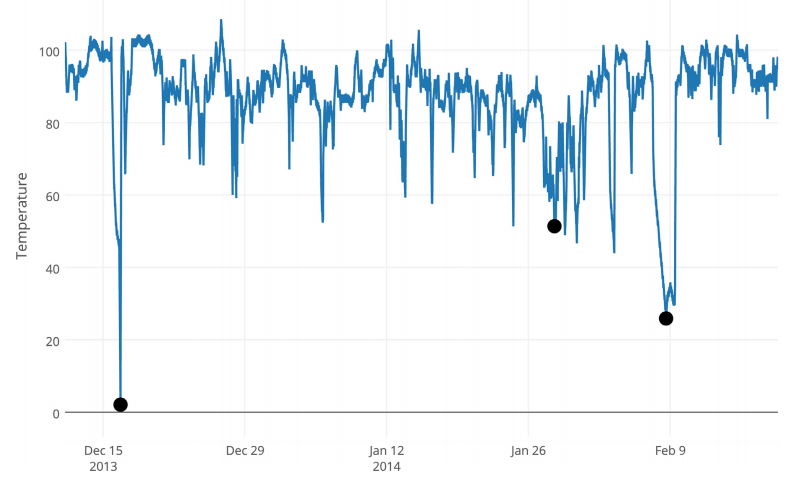


Figure - Real world Temperature Sensor Data with Anomalies marked as Dark circles.

The authors characterize the ideal qualities of a certifiable peculiarity location calculation as follows: Forecasts should be made on the web; i.e., the calculation should recognize state ‘*xt’* as ordinary or irregular prior to getting the ensuing xt+1.

1. The calculation should adapt ceaselessly without a necessity to store the whole stream.
2. The calculation should run in an unaided, mechanized style—i.e., without information names or manual boundary tweaking.
3. Calculations should adjust to dynamic conditions and idea float, as the fundamental measurements of the information stream is frequently non-fixed.
4. Calculations should make peculiarity discoveries as ahead of schedule as could really be expected.
5. Calculations ought to limit bogus positives and bogus negatives (this is valid for group situations too).

Taken together, the above prerequisites propose that peculiarity location for streaming applications is an in a general sense unique issue than static cluster anomaly detection.

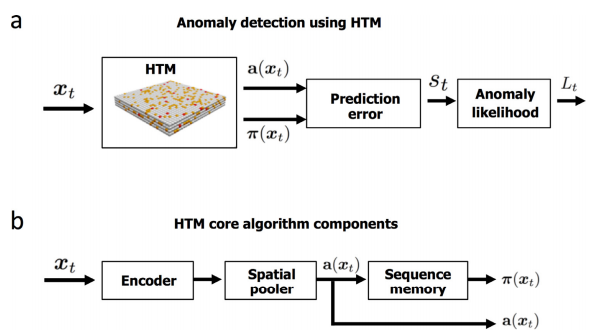


Figure - Methodology to Create an Anomaly Detector based on HTM

The contributions of this paper are twofold: a novel irregularity discovery procedure worked for continuous applications, and a far-reaching set of results on a benchmark intended for assessing inconsistency discovery calculations on streaming information.

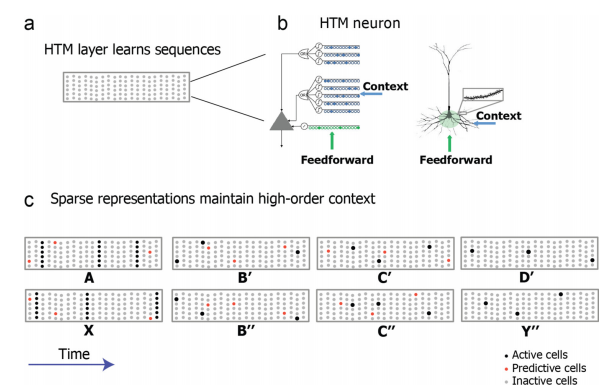


Figure - HTM (Hierarchical Temporal Memory Workflow)

Given the current information, ‘*xt, a(xt)’* is a scanty encoding of the current information, and π(xt−1) is the inadequate vector addressing the HTM organization's interior expectation of *‘a(xt)’*. The dimensionality of both vectors is equivalent to the quantity of sections in the HTM organization (we utilize a standard estimation of 2048 for the quantity of segments in every one of our analyses). Let the forecast mistake, *‘st’*, be a scalar worth conversely relative to the quantity of pieces basic between the real and anticipated paired vectors:

; where |a(xt)| is the scalar standard, for example the absolute number of 1 piece in a(xt). The mistake st will be 0 if the current a(xt) impeccably matches the forecast, and 1 if the two paired vectors are symmetrical (for example they share no basic 1 pieces). this condition gives us a quick proportion of how well the basic HTM model predicts the current info xt.

The aim of the NAB dataset is to present algorithms with the challenges they will face in real-world scenarios, such as a mix of spatial and temporal anomalies, clean and noisy data, and data streams where the statistics evolve over time. The best way to do this is to provide data streams from real-world use cases, and from a variety of domains and applications. The data currently in the NAB corpus represents a variety of sources, ranging from server network utilization to temperature sensors on industrial machines to social media chatter.

NAB version 1.0 contains 58 data streams, each with 1000 - 22,000 records, for a total of 365,551 data points. Also included are some artificially-generated data files that test anomalous behaviours not yet represented in the corpus’s real data, as well as several data files without any anomalies. All data files are labelled, either because we know the root cause for the anomalies from the provider of the data, or as a result of the well-defined NAB labelling procedure. These labels define the ground truth anomalies used in the NAB scoring process.

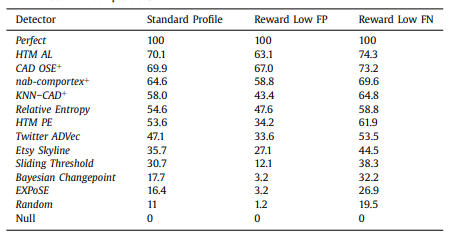


Figure - Results of Different Algorithms with NAB Dataset

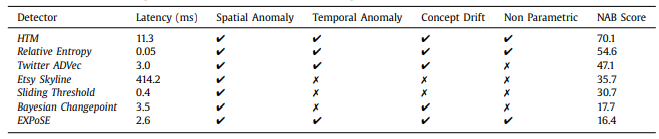


Figure - Properties Comparison for Different Algorithms on the NAB Dataset

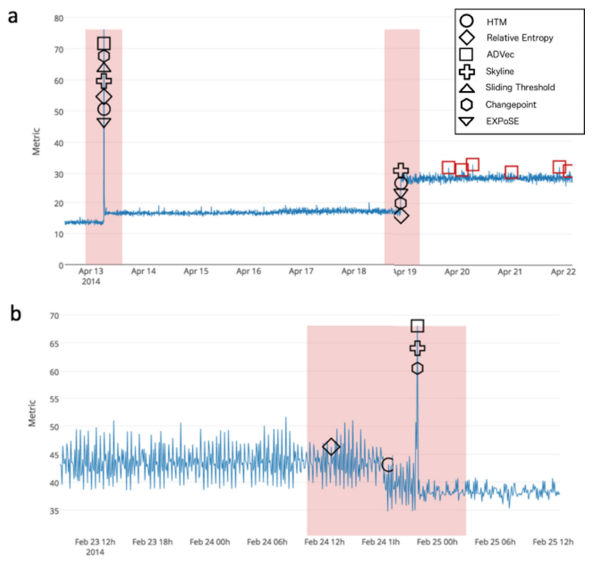


Figure - Real-time Detection Results of Different Algorithms on the NAB Dataset

### A Comparative Evaluation of Unsupervised Anomaly Detection Algorithms for Multivariate Data

in this examination, 19 diverse solo Anomaly Detection calculations are assessed on 10 diverse datasets from various application areas. By distributing the source code furthermore, the datasets, this paper intends to be another very much financed reason for solo abnormality location research. Also, this assessment uncovers the qualities and shortcomings of the various methodologies interestingly. Other than the irregularity recognition execution, computational exertion, the effect of boundary settings just as the worldwide/nearby irregularity identification conduct is laid out. As an end, we give a prompt on calculation choice for commonplace certifiable undertakings.

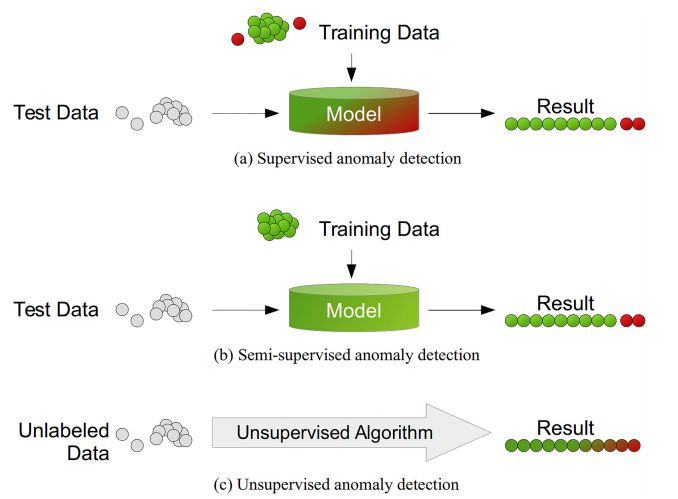


Figure - Different Anomaly Detection Methodologies Based on the Machine Learning Style

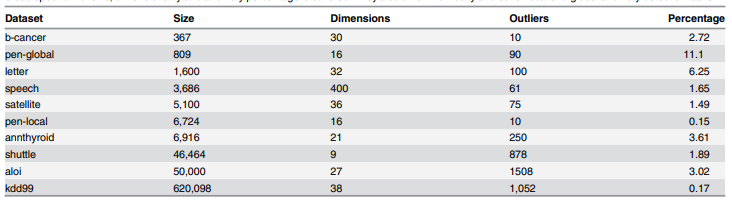


Figure - Datasets used for comparison in this Research

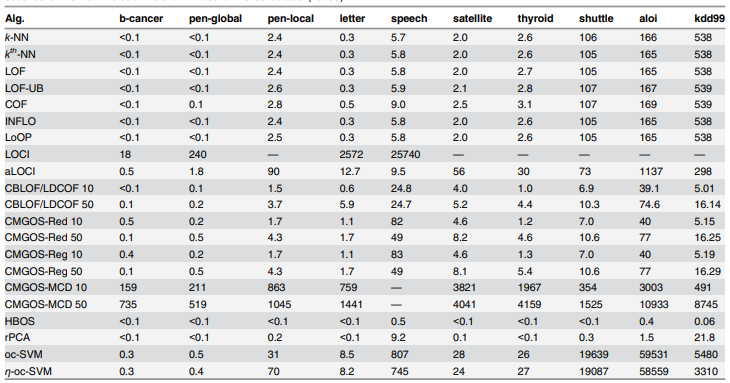


Figure - Results for All the Algorithms for the Datasets used.

## Project Timeline

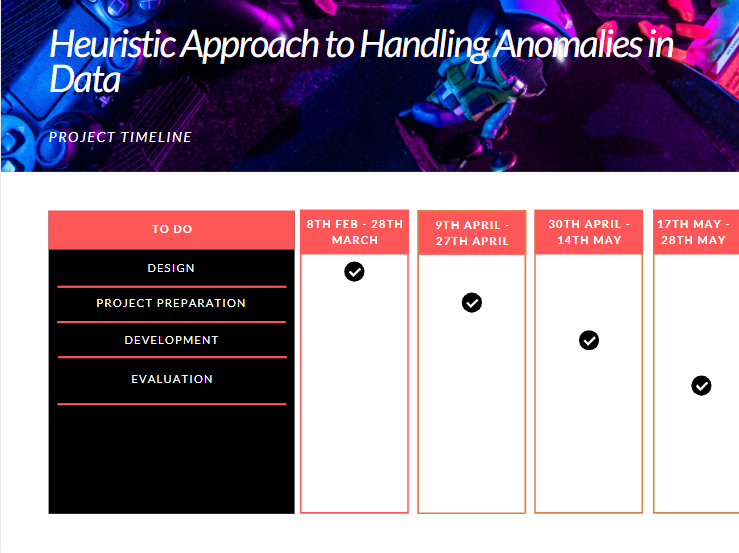


Figure 12 - Project Timeline

# Methodology

The Literature Review Performed for this research covers only the classification approach for this domain. This research is focused on the implementation of Regression-based algorithms for targeting the missing data anomalies in datasets. There are plenty of toy datasets available for this purpose of training the algorithms to fill the gaps in the dataset using regressive training. There can be multiple datasets used to evaluate multiple regression approaches. Further research will allow a final decision on this.

# Project Evaluation

The project evaluation will be done using cross validation. The dataset will be divided into *n* parts. Each part will contain missing values. The missing values will be stored for comparison in another dataset. The algorithm will be asked to predict the missing values. The standard deviation between both the predicted and the original values will evaluate the algorithm’s performance.

# Conclusion

As the importance of data increases exponentially with the advent of technology, researchers move towards faster and more efficient ways of collecting data. This resulted in the automation of the former process. But automation isn’t perfect due to the absence of the human element. There are issues with the dataset collected. In most cases it has missing values. These missing values cause a drop in the quality of the dataset, which in-turn drop the quality of the Machine Learning algorithm’s predictions. This research focuses on using Regression to find a solution to this problem.

# References

* Goldstein, M. and Uchida, S., 2016. A Comparative Evaluation of Unsupervised Anomaly Detection Algorithms for Multivariate Data. PLOS ONE, 11(4), p.e0152173.
* Agrawal, S. and Agrawal, J., 2015. Survey on Anomaly Detection using Data Mining Techniques. Procedia Computer Science, 60, pp.708-713.
* Ahmad, S., Lavin, A., Purdy, S. and Agha, Z., 2017. Unsupervised real-time anomaly detection for streaming data. Neurocomputing, 262, pp.134-147.