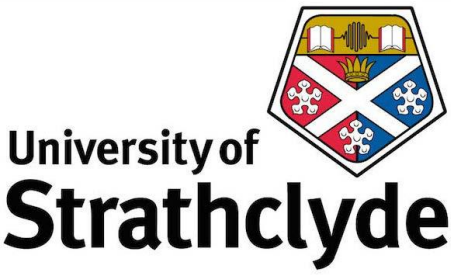
# Location Fingerprinting for IoT Systems

# Using Machine Learning

**Ibrahim Aqeel**




**Doctor of Philosophy**

**2020**

# Location Fingerprinting for IoT System Using Machine Learning

Ibrahim Aqeel, 2020

University of Strathclyde

Electronic & Electrical Engineering

Royal College Building

204 George St

Glasgow G1 1XW

SCOTLAND

# ACKNOWLEDGEMENTS

I would like to acknowledge everyone who provided support and guidance throughout my studies.

Firstly, my everlasting gratitude to my Family for their love and understanding; my Mother's soul, my Father, my wife, my children (Ahmed, Abdul-Malik and my beautiful little girl Mlath), sisters, mother-in-law, Anwar, Murad and Mohannad. Too many years away from you, but without you, I could never have reached my goal.

Secondly, my supervisors, each of whom has provided patient advice and guidance throughout. I would like to express my sincere gratitude to my advisor Prof. Ivan Andonovic for the continuous support of my PhD study and related research, for his patience, motivation, and immense knowledge. He has treated me as a member of his family since I first met him, and I never really felt that I was an expatriate. In addition, I would like to thank Dr Christos Tachtatzis and Dr Robert Atkinson for their insightful comments and encouragement, but also for the hard challenges that widened the scope of my research. Thank you for your unlimited generosity. Thanks also to my colleagues Hussain Zangoti, Mohammad Abohalalh and Ephraim Iorkyase for advice.

My thanks also to the Emirate of Jazan Province which facilitated the ready access to the environment within Jazan City and to the Jazan Municipality and the Ministry of Environment Water and Agriculture for making available its resources that enabled the data acquisition, core to the research developed. I am indebted to Abdul-Rahman Saheli, Majed Ismail and Issam Al-Hamzi for granting the necessary permits.

Last but not least, my sincere gratitude to my country, the Kingdom of Saudi Arabia, which has provided generous financial support that have enabled my PhD studies.

## إهداء

الحمد الله رب العالمين و الصلاة و السلام على أكرم الخلق نبينا محمد عليه افضل الصلاة و اتم التسليم،

بحمد لله و توفيقه اتممت كتابة رسالة الدكتوراه و التي أسأل الله لها القبول، و التي أحببت أن أهديها

أولاً لعائلتي:

لروح أمي الطاهرة ملهمتي الدائمة. لوالدي الذي لم أكن لأجتاز هذه المرحلة لولا الله ثم دعمه اللامحدود سواءً مادياً أو معنوياً فشكراً من القلب و جعلك الله ذخراً لي و أدام الله عليك صحتك و عافيتك.كما أهديها الى زوجتي الحبيبه و الى أبنائي أحمد،عبدالملك و صغيرتي ملاذ، شكراً لتحملكم المسؤولية و استيعابكم للظروف فكنتم نعم العون لي. ولا انسى دعم أخواتي وعمتي و جميع أهلي حفظهم الله اللامحدود فالشكر و الإمتنان موصول لكم. الشكر ايضاً لمراد شامي و العم أنور عقيل على دعمهم و تخفيفهم معاناة البعد و السؤال المتواصل. و اخيراُ صديقي و اخي مهند زرير، كل يوم من ايام ابتعاثي ستجد فيه بصمة من بصماتك الخيرة،كل صفحة من هذه الرسالة و كنت لي خير المعين فيها، هذه الرسالة هي هديتي لك.

ثانياً أهديها لمشرفيني:

البرفسور آيفن اندونوفيك، مشرفي الرئيسي، الاسكتلندي البشوش صاحب الروح الأبوية، منذ لقائي الأول معه و لم أشعر بأني مغترب. كان تعامله معي كأحد افراد عائلته. كان حريص أن يصقل مهاراتي العملية و العلمية، كان حريصاً أن أتعلم كيف ابني البيانات و كيف اجمعها ثم احللها بالطريقة العلمية. لم يكن هذا العمل ان يُنجز لولا مرونته خاصتاً في دعمي في رحلاتي العلمية و متابعته الدائمه لأداء البحث. أود ان اشكر ايضاً مشرفي الثاني الدكتور روبرت اكتنسون على دعمه و تزويدي بالمصادر في السنة الأولى لتوسيع مداركي المعرفيه. كما أشكر ايضاً المشرف الدكتور كرستيوس تكتاتزايس، في كل اجتماع او نقاش معه في المكتب او حتى في اروقة الكلية وان كانت الفترة قصيرة الا انها تضيئ لي أيام من العتمة. كان كرمه لا محدود معي، اغلب المصادر كان يزودني بها لتثري بحثي، شكرا دكتور كرستيوس على هذا الدعم. شكراً للصديق و الزميل حسين زنقوطي على دعمي في مراحل البحث الأولى. و لا أنسى أن اشكر الدكتور محمد آل سالم على توجيهاته القيمة و الثمينة و حرصه ان يكون البحث داخل المملكة العربية السعودية لهدف الاستفادة من الدراسة مستقبلاً.

ثالثاً أهديها للجهات التي استضافت و دعمت البحث الميداني:

1- أمارة منطقة جازان، ممثلة بأميرها صاحب السمو الملكي الأمير محمد بن ناصرين عبدالعزيز آل سعود. و نائب أمير المنطقة الامير محمد بن عبدالعزيز بن محمد بن عبدالعزيز آل سعود، على تبنيهم لهذا العمل و تسهيل جميع الإجراءات و التصاريح اللازمة،فشكراً لكم جميعاً على هذا الدعم.
2- أمانة منطقة جازان مستضيف البحث، اشكر لكم استضافة بحثي و دعمه مادياً و معنوياً. شكراً لسعادة الأمين السابق: م.عبدالله الدبيان، شكراً لسعادة الأمين الحالي الأستاذ نايف بن سعيدان، و شكراً للانسان الخلوق الداعم في كل مرحلة من مراحل البحث م.عبدالرحمن ساحلي. ولا انسى الاستاذ ماجد اسماعيل على حرصة بتوفير كل الاحتياجات الاساسية لهذا البحث.
3- وزارة البيئة و المياه و الزراعة فرع منطقة جازان ممثلة بمديرها و بالأستاذ عصام الحمزي على حسن التعامل و السماح لنا باستخدام أدواتهم في البحث الميداني.

أخيراً وليس آخراً، أهدي هذا العمل لوطني الحبيب المملكة العربية السعودية التي لم تبخل علي بشيئ من دعم مادي أو معنوي، فخري لا حدود له بإنتمائي لهذا البلد العظيم.

حفظ الله بلادي و حفظكم و أدام الله أمننا و أماننا

إبراهيم بن أحمد إبرهيم عقيل

# DECLARATION

I declare that this Thesis has been composed solely by Ibrahim Aqeel and that it has not been submitted, in whole or in part, in any previous application for a degree. Except where stated otherwise by reference or acknowledgement, the work presented is entirely my own.

Signature: 

Date: 26th March 2020

# ABSTRACT

The Internet of Things (IoT) has evolved rapidly as the number of connected nodes continues to grow, projected to be in excess of Trillions worldwide by 2025. IoT enables a number of application and services that enhance the quality of life of citizens and business practice. The demand for IoT-like connectivity is set to continue; for example, the advent of LPWAN providing a combination of advantageous features such as long-range, low power connectivity gates the deployments of a range of hitherto costly implementations over extended areas of coverage.

A spectrum of valuable real-world IoT applications such as tracking, are predicated on location information. However, the provision of a low power, cost effective engineered solution to provisioning location still remains a major challenge, especially within resource constrained IoT deployments. GPS-enabled solutions are power hungry and potentially prohibitively expensive within extensive IoT architectures. Furthermore, ranging-based network-centric methods lack accuracy because of the long distances subject to dynamically varying path characteristics and the ultra-narrow bandwidth. The prevailing state-of-the-art motivates investigations into low-complexity, energy-efficient technique for IoT node localisation.

The Thesis presents an empirical investigation into the use of fingerprinting for IoT node localisation within a suburban region in Saudi Arabia subject to varying environmental conditions, ranging from clear sky to sandstorms. The approach is based on the use of Received Signal Strength Indicator (RSSI) within a LoRaWAN network setting.

The performance of LoRa transmission as a function of varying coding parameters is determined. The RSSI data gathered during the characterisation phase is exploited to estimate locations of IoT nodes using location fingerprinting. More specifically, k-Nearest Neighbour (KNN) algorithms are used to develop a baseline location model.

The accuracy of the LoRaWAN based baseline node localisation is enhanced through the use of Machine Learning (ML). RSSI ratios between pairs of Gateways in conjunction with kernel-based ML techniques - Support Vector Regression (SVR) and Gaussian Process Regression (GPR) – is proven to improve the node localisation models. Moreover, the impact of the kernel function on model performance is evaluated. Further, RSSI measurements at different spreading factors are combined to form more robust location features; two machine learning ensemble techniques - Gradient Boosting and Random Forest - are then employed to determine the impact on the accuracy of node localisation models using combined location features. Results indicate that ensemble-derived models improve accuracy compared to single regression tree methods. In addition, feature transformation is proven to be effective in improving localisation performance.

Results confirm the feasibility of IoT network-derived localisation in sandstorm environments. Furthermore, it is demonstrated that the LoRaWAN spreading factor is central to optimising performance.

# CONTENTS

# FIGURES

# TABLES

# ACRONYMS

| | |
|---|---|
| **ADR** | **Adaptive Data Rate** |
| **A-GPS** | **Assisted Global Positioning System** |
| **AoA** | **Angle of Arrival** |
| **BW** | **Bandwidth** |
| **CDF** | **Cumulative Distribution Function** |
| **CSS** | **Chirp Spread Spectrum** |
| **CR** | **Code Rate** |
| **CRC** | **Cyclic Redundancy Check** |
| $d$ | **Distance** |
| **EM** | **Electro Magnetic** |
| **FEC** | **Forward Error Correction** |
| **FHSS** | **Frequency-Hopping Spread Spectrum** |
| **FSK** | **Frequency-Shift Keying** |
| **GBR** | **Gradient Boosting Regression** |
| **GNSS** | **Global Navigation Satellite System** |
| **GLONASS** | **Global Navigation Satellite System** |
| **GPS** | **Global Positioning System** |

| | |
|---|---|
| **GPR** | **Gaussian Process Regression** |
| **GSM** | **Global System for Mobile Communication** |
| **GW** | **Gateway** |
| **IoT** | **Internet of Things** |
| **KNN** | **K-Nearest Neighbour** |
| **LoRaWAN** | **Long Range Wide Area Network** |
| **LPWAN** | **Low Power Wide Area Network** |
| $L$ | **loss function** |
| **LDPs** | **Location Dependent Parameters** |
| **LoS** | **Line-of-Sight** |
| **LF** | **Location Fingerprinting** |
| $lng, lat$ | **Longitude and Latitude** |
| **M2M** | **Machine-to-Machine** |
| **MSC** | **Mobile Switching Centre** |
| **NLOS** | **Non-Line-of-Sight** |
| **PER** | **Packet Error Rate** |
| **PL** | **Path Loss** |
| **RSSI** | **Received Signal Strength Indicator** |
| **RDT** | **Regression Decision Tree** |

| | |
|---|---|
| **RFR** | **Random Forest Regression** |
| **$R_c$** | **chip rate** |
| **$R_S$** | **symbol rate** |
| **$R_b$** | **bit ate** |
| **RPi** | **Raspberry Pi** |
| **SF** | **Spreading Factors** |
| **SVR** | **Support Vector Regression** |
| **ToA** | **Time of Arrival** |
| **TDoA** | **Time Difference of Arrival** |
| **TTN** | **The Things Network** |
| **$T_S$** | **Symbol duration** |
| **WKNN** | **weighted K-Nearest Neighbour** |

# CHAPTER 1 INTRODUCTION

## 1.1 Introduction

Recent advances in highly functional sensor nodes and wireless communication standards interfaces have resulted in the ready availability of a range of low-cost, high performance elements. A combination of such technologies integrated into systems classified as the ''Internet of Things (IoT)" have, as a consequence, evolved at a rapid rate yielding demonstrable impact in many application scenarios throughout the world.

IoT is characterised by highly inter-connected distributed networks of 'things' - whether physical or electronic - across environments at different locations, collaborating to execute on an application or monitor processes or citizens. IoT technologies, systems and applications have been subject to a significant amount of research and development and IoT-inspired solutions are currently experiencing increased levels of adoption. Continual technology developments further stimulate the evolution of the discipline, allowing improved performance and efficiency of execution. Interconnected nodes can exchange data and information from different locations [1] and can be controlled and monitored remotely, not necessarily requiring human intervention. The degree of inter-connection is continually increasing as technology progresses, the projection being, that by 2025, the number of IoT connected devices will rise to 75 billion translating to a net addressable market opportunity of >$11trillion per annum [2] [3]. Such potential is a significant driver in accelerating the transformation of the quality-of-life of citizens, management of the environment and industrial practices. The ability to automate certain tasks through the utilization of IoT-

generated data coupled with the context of the task/goal/service better informs decision making.

However, the further enhancement of the functionality of IoT implementations still presents a number of challenges. Location is one of the critical features with accuracy of estimation remaining a key challenge, if solved, unlocks many more applications in health monitoring, transportation, intrusion detection and environmental monitoring. IoT harnesses many communications technologies, such as WiFi, Bluetooth, ZigBee, infrared, GPRS, and more recently 5G, which not only provide robust connections between different entities but gate a number of mechanisms to obtain location information.

A range of wireless sensor network localisation techniques have been developed for both indoor and outdoor environments. However, although existing location estimation techniques based on the Global Positioning System (GPS) achieve the desired accuracy with a 3 – 5 meters [4], their use is not only expensive due to a relatively sophisticated infrastructure but consume notable levels of energy. GPS-derived location information is thus limiting within the scope of IoT implementations, useful as a complementary solution but not in most large scale deployments.

Other network-based methodologies for deriving location can be broadly categorised into range-based and range-free localisation [5] [6]. Range-based localisation consists of ranging with location computation. In the first phase, methods such as Time of Arrival (ToA) [7], Time Difference of Arrival (TDoA) [8], Angle of Arrival (AoA) [9], and Received Signal Strength Indicator (RSSI) [10] are utilised for ranging to obtain the distance between two

nodes. In the second phase, trilateration, triangulation or maximum likelihood estimation are used together with coordinate information of reference nodes and RSSI to estimate the location of any node in the network.

Time based localisation methods (ToA and TDoA) require extra hardware to guarantee time synchronisation between transmitter and receiver; any small timing error may result in a large distance estimation and in turn, location error. Angle-based localisation (AoA) allows the estimation of distance according to relative angles, accomplished by suitable measurement equipment; the latter increases the cost of large-scale deployments. The Received Signal Strength Indicator (RSSI), on the other hand, does not require any additional hardware and utilises signal propagation models to translate a signal strength to distance. However, the technique suffers from multi-path propagation that compromise the accuracy of the ranging estimate. Moreover, the signal propagation models are generic and do not capture the complexities of the operational environment. These challenges have motivated further investigations into the use of the RSSI-based fingerprinting techniques for more accurate and cost-effective IoT node localisation. The approach can in addition, harness machine learning to capture the characteristics of the network connection patterns through a systematic range of RSSI measurements from different locations with the resultant model used to infer node location.

## 1.2 Research Objectives

The need for an engineered network-derived node localisation for extensive IoT implementations in suburban environments subject to sandstorm conditions in Saudi Arabia

is the motivation underpinning the research. The proposed solution is based on the readily attainable Received Signal Strength Indicator (RSSI) acquired within a network of LoRaWAN nodes, enhanced through fingerprinting and machine learning. The localisation solution has the potential to gate an increased number of IoT applications and services as will proving the feasibility of an acceptable performance within the challenging operational environment under investigation.

Specifically, the objectives of the research are to;

- carry out a review on localisation techniques for IoT systems
- develop a RSSI-based fingerprinting technique for node location estimation within extensive IoT deployments
- investigate the effect of LoRaWAN parameters e.g. spreading factors on localisation performance
- investigate feature transformation techniques for enhanced localisation
- develop localisation models using machine learning
- to confirm the feasibility of obtaining IoT network-derived localisation in sandstorm environments

### 1.3 Main Contributions

The research reported contributes to knowledge in the following ways;

- development and characterization of a LoRaWAN based IoT node localisation method using fingerprinting with machine learning within a sandstorm (suburban)

environment. The solution demonstrates the attainable performance of RSSI-based localisation of IoT deployments in challenging environments through a limited number of Gateways and significant distances between nodes.

- evaluation of the effect of spreading factors on the performance of localisation with fixed nodes.
- engineering of new location features to improved node localisation; the improvement in using the ratios of RSSI received by Gateways and a combination of RSSIs for different spreading factors are evaluated.
- development of new IoT node localisation models for sandstorm environments based on fingerprinting and machine learning techniques. Results are compared with reported performance and provide evidence that an acceptable accuracy in locating nodes in such a challenging environment can be obtained.

## 1.4 Thesis Structure

The Thesis comprises seven Chapters. Chapter One is an introduction to the research, the aims and objectives of the study, capturing the main contributions to knowledge and summarizing the outline of the Thesis.

Chapter Two is a review of related literature with emphasis on localisation methods within IoT, LPWAN and the application of machine learning to node location.

In Chapter Three, the operational environment and the methodology for data collection using LoRaWAN is presented. Two data gathering phases executed are detailed; the first conducted during clear weather whilst the second during severe sandstorm periods. The two

phases were necessary to capture the effect of sand storms on radio propagation. Both the LoRaWAN nodes and Gateways are deployed at arbitrary but known locations in the chosen environment, a suburban area of the City of Jazan in Saudi Arabia.

Chapter Four details an investigation into the use of RSSI-based fingerprinting for LoRaWAN based radio-location of IoT nodes in the chosen environment. Node location is determined by using RSSI as input to a localisation algorithm. K-Nearest Neighbour and its variant, weighted k-Nearest Neighbour algorithms are used as a baseline for the development of the node location models; a performance analysis of the developed location models is presented.

Chapter Five discusses the implementation of kernel-based node localisation. More specifically, two kernel-based techniques, Support Vector Regression and Gaussian Process Regression are used to establish location models. Moreover, robust location features are computed and used as input to the models. The impact of the kernel function is investigated through a performance evaluation.

Chapter Six focuses on methods to improve the accuracy of node location. Feature combination and ensemble machine learning techniques are explored. In feature combination, the RSSI values of different spreading factors are combined to form new location features used to infer node location. Furthermore, tree based machine learning ensemble methods viz. Random Forest and Gradient Boosting formulated as a regression problem are investigated and used to model the complex relationship between the features and node location; the models are then used to infer sensor location. A performance analysis of the techniques used to enhance node localisation is presented.

Chapter Seven is a summary of the findings of the research drawing conclusions on the feasibility of the approach and providing recommendations for future developments.

# CHAPTER 2 REVIEW OF IOT NODE LOCATION ESTIMATION

## 2.1 Introduction

The proliferation of inexpensive sensor nodes and wireless communication technologies with enhanced performance has enabled the rapid deployment of significant IoT applications and services for both citizens and industries, paving the way for the transformation of current practices [11] [12] [13]. Stimulated by the benefits owing to IoT solutions, the location of nodes or entities has become increasingly important in order to further increase the functionalities and application environments of such implementations [14] [15]. Several techniques have been developed in the goal of solving the challenges in obtaining network-enabled localisation based on signal strength or angle, in some cases enhanced in performance through machine learning techniques.

The Chapter provides a review of the state-of-the-art in IoT node location methods beginning with a general overview of wireless technologies (as it relates to IoT) and concluding with a review of reported localisation techniques. A case is made for one particular wireless standard - LoRaWAN – for the implementation of a localisation method applicable to extensive IoT deployments.

## 2.2 Wireless Technologies for IoT

One of the important considerations for many IoT network implementations is the selection of the most appropriate communication technology that meets the requirements of the application at the lowest cost (Figure 2.1). While some IoT network applications require

short range (meters) (Bluetooth, ZigBee and NFC) [10] [16] others need long-range radio (kilometers); the suitability of a long-range communication technology varies with application. For example, 2G, 3G, 4G and LTE technologies support long-range operation but consume significant energy, however, Low Power Wide Area Network (LPWAN) radio presents a suitable candidate for IoT applications meeting the prime design requirements of low power consumption coupled to low cost [17] [18].

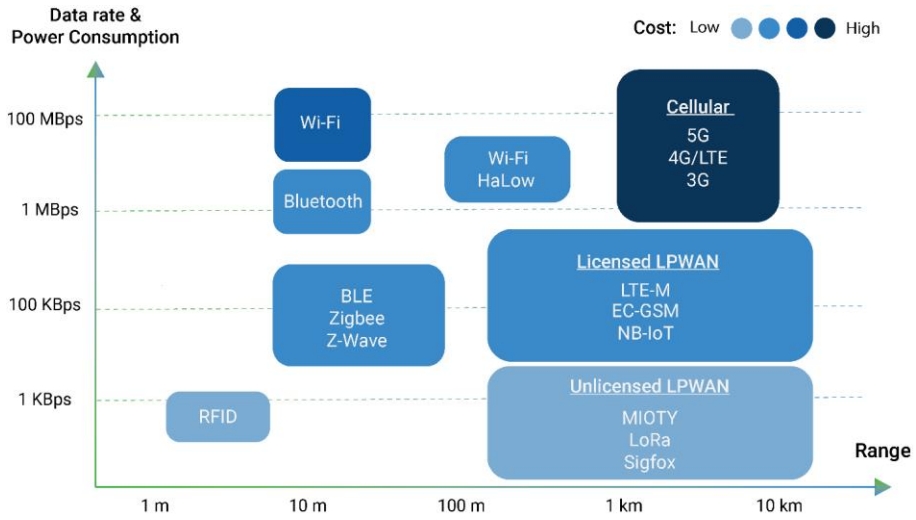



**Figure 2.1: Communication technology for IoT [19].**

Several LPWAN network standards have been designed to provide an appropriate solution to many deployment issues including Sigfox, LoRaWAN, NB-IoT, LTE-M and Weightless [20] [21] [22] [23]. All have their advantages and limitations, in terms of cost, power consumption, data rate, scalability etc. For applications requiring the transport of modest data rates, most wireless technologies are applicable. However, LoRaWAN offers the highest radio link budget and the best "cost vs. range vs. power" trade-off [24] [25]. Therefore, the Thesis adopts the LoRaWAN network standard to implement the proposed node localisation method for sandstorm environments.

## 2.3 LoRaWAN

Long Range Wide Area Network (LoRaWAN) is a low-power radio that enables long range, low power consumption networking at the expense of low data bandwidth (Figure 2.2). LoRa is thus suitable for networks that require longer range communications with feature constraints on the size of the nodes, power consumption and cost [26] [27] [28].

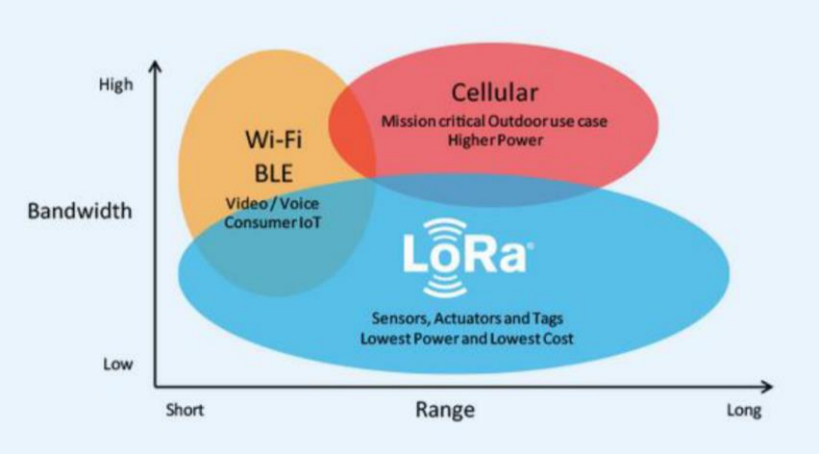



**Figure 2.2: LoRa in comparison with other wireless technologies [29].**

The LoRaWAN-layer structure is as shown in Figure 2.3, established upon the LoRa modulation scheme with an added network layer to manage data traffic between Base (central) Stations and end-user nodes [26]. The standard has been developed specifically for extensive IoT implementations, wide-area sensor networks, and machine-to-machine (M2M) applications. The radio interface is designed to enable extremely low signal levels to be received, extending the transmission path distance significantly. LoRa is the first commercially available wireless technology with the combined advantages of low cost, long transmission range, and low power consumption [30]. More specifically, LoRa has a stated line-of-sight range of up to 15km-36km, a data rate capability of up to 50kbps, and a 10-year

battery life [31] [32]. Although intended for use outdoors, its functionality can also be utilised for indoor applications [33]; given that LoRa operates in the sub-GHz frequencies, it has a greater penetration ability and is, thus, more resilient to noise and multi-path interference [34] .

One of the added properties of LoRa is its ability to provide the foundation for the estimation of node location extending its use for example, within large facilities (e.g. multistory buildings, large warehouse) where multiple access points required to execute on the feature due to the short range of traditional wireless standards e.g. Bluetooth. With the extended range of LoRa, fewer access points can sustain similar operations [35].

In conclusion, LoRa offers an efficient, flexible and economical solution to real-world problems in outdoor and indoor use cases, overcoming the limitations of some other wireless based networks [24] [30].

Therefore, the focus of the research is the development and performance evaluation of an extensive IoT network localisation technique using LoRa technology.

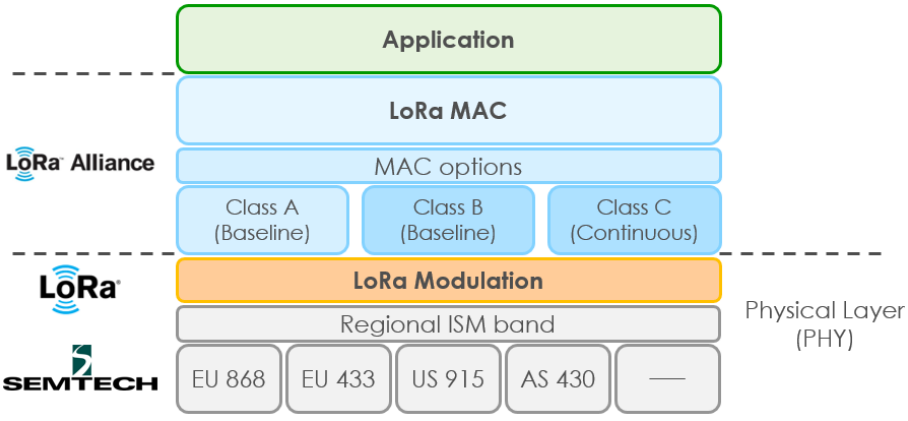



**Figure 2.3 Layer structure of LoRaWAN [36].**

### *2.3.1 Network Architecture*

LoRaWAN is an open standard developed by the LoRa Alliance that enables numerous nodes ("End devices") to communicate with receivers ("Gateways") using the LoRa modulation [37]. A typical LoRa network consists of three entities: Gateways, nodes, and a network server as shown in Figure 2.4.

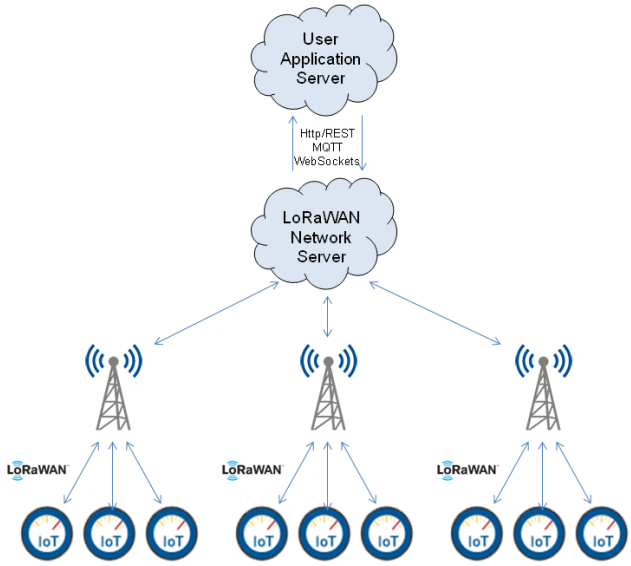



**Figure 2.4 LoRaWAN network architecture [38].**

Typically, but not confined to, nodes may comprise sensors generating measurement data connected to the Gateways, where the latter is regarded as the 'central node'. Nodes communicate with Gateways that then forward raw LoRaWAN frames to a network server over an interface with a greater throughput. Gateways are bi-directional [25] relays transferring packets to the server where they are decoded, or from the server where they are encoded, to the nodes. The network architecture is a "star-of-stars" topology with a server as a central node and the Gateways intermediate nodes. Further, the nodes may form

part of the Internet of Things (IoT) where a network containing a variety of objects or "things" are connected to the Internet and are configured to provision a service or monitor an asset [39].

### *2.3.2 LoRa Physical Layer Parameters*

A key factor in implementing LoRaWAN networks is the modulation format, a Chirp Spread Spectrum (CSS) scheme which creates a wideband linear frequency variation over time in order to encode the transmitted information [40]. A chirp is generated by modulating the phase of the signal. An "up-chirp" as shown in Figure 2.5 progresses from a minimum to maximum frequency before wrapping around to the minimum; a "down-chirp" is the reverse.

Data is encoded for transmission through discontinuities in individual chirps, as shown in the CSS transmission representation of Figure 2.6. Each modulation is referred to as a 'chip' and the number of times per second the phase is modulated is referred to as the chip rate. The chip rate is equal to the spectral bandwidth of the signal which occupies a bandwidth of 125 kHz, 250 kHz or 500 kHz. Given that the chirp spreads the spectrum, the entire bandwidth is utilised in the transmission of a signal, increasing the robustness to channel noise [41] and multi-path fading [20]. The time-varying frequency of the chirps also minimises the effect of the Doppler spread on the channel.

High precision chirps are generated using inexpensive crystals leading to low chip cost. Furthermore, LoRa employs a Frequency-Hopping Spread Spectrum (FHSS) scheme to switch frequency between available channels determined by a pseudo-random distribution, helping further to mitigate against interference. The features outperform traditional

modulation schemes such as Frequency-Shift Keying (FSK), and makes LoRa particularly suited to low-power, long-range applications. LoRa provides a line-of-sight transmission range of 30km and a 15km range in non-line-of-sight scenarios [31].

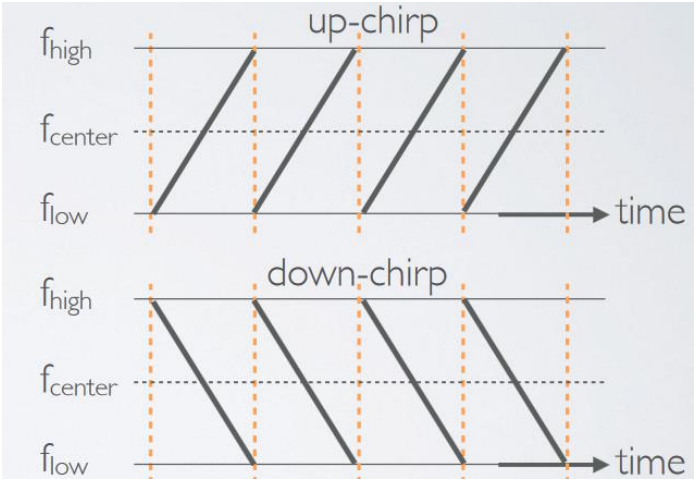



**Figure 2.5: Up- and down-chirp [42].**

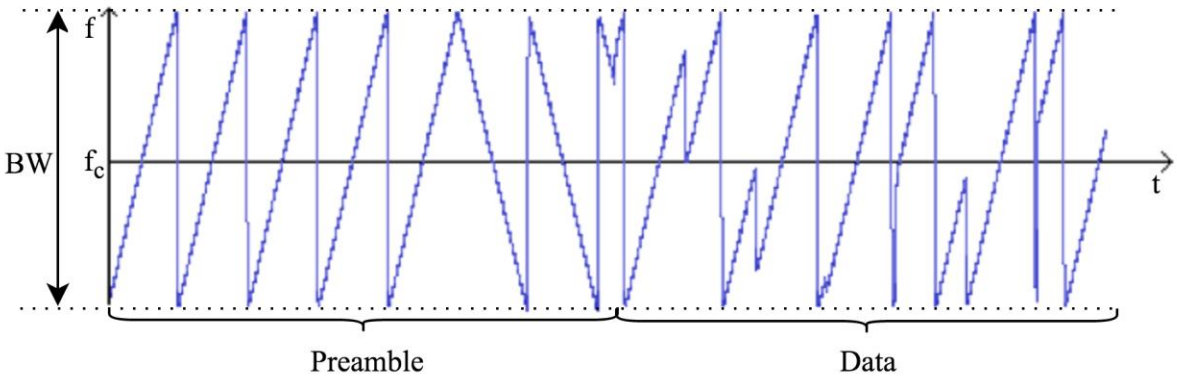



**Figure 2.6: A typical chirp spread spectrum transmission [41].**

There are three configurable parameters in LoRa modulation: Spreading Factor (SF), Bandwidth (BW), and Code Rate (CR). The spreading factor is defined as the logarithm, in base 2, of the number of chips per chirp (or "symbol"). A symbol comprises $2^{SF}$ chips over the full bandwidth. Given that there are $2^{SF}$ chips per symbol, a symbol can effectively encode

SF bits of information. In turn, SF impacts the time of a transmission as it spreads the signal over time; hence spreading factor. The SF value can be set between 7 and 12 in increments of 1, where each increment approximately doubles the time taken to transmit the signal [43].

The bandwidth determines the chip rate ($R_c$), which is equal to one chip per second per hertz of bandwidth. Consequently, an increase of one in SF divides the frequency span of a chip by two; thus the duration of a symbol is multiplied by two, as shown in Figure 2.7.

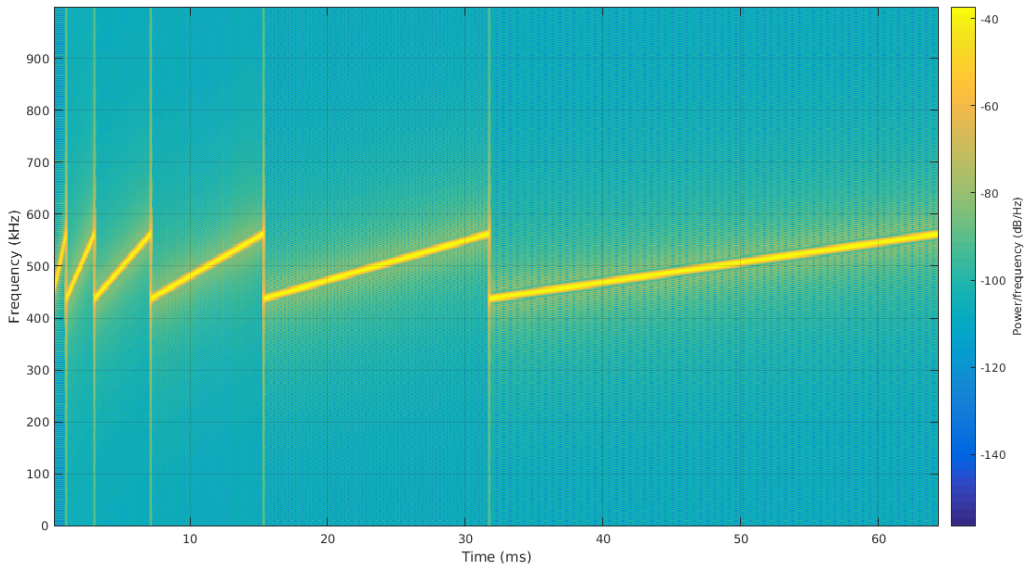



**Figure 2.7: Comparison of LoRa spreading factors [44].**

The symbol rate and the bit rate are proportional to the bandwidth; therefore, an increase in the bandwidth produces a corresponding increase in the transmission rate. Equation 2.1 governs the relationship between symbol rate ($R_S$), bandwidth (BW) and the spreading factor (SF) [45];

$$R_S = \frac{BW}{2^{SF}} \qquad \textbf{Equation 2.1}$$

LoRa features a forward error correction code (FEC); the code rate (CR) equals $4/(4 + n)$, where $n \in \{1, 2, 3, 4\}$. Taking the CR into account and that SF bits of information are transmitted per symbol, the useful bit rate ($R_b$) is given as in Equation 2.2;

$$R_b = SF \times \frac{BW}{2^{SF}} \times CR \qquad \textbf{Equation 2.2}$$

An increase in BW increases the bit rate whereas an increase in SF decreases the bit rate or data rate. These parameters also affect the sensitivity of the receiver. An increase in bandwidth lowers receiver sensitivity, whilst an increase in SF increases receiver sensitivity. A lower CR reduces the Packet Error Rate (PER) during which there are shorter bursts of interference [41].

Table 2.1 shows the receiver sensitivities for a Semtech SX1276. LoRaWAN provides an Adaptive Data Rate (ADR) mechanism which utilises different SFs and bandwidths for managing data rates, airtime, and energy consumption within the network.

**Table 2.1: RF sensitivity depending on BW and SF [46].**

| BW\SF | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|
| **125kHz** | -123dBm | -126dBm | -132dBm | -132dBm | -133dBm | -136dBm |
| **250kHz** | -120dBm | -120dBm | -128dBm | -128dBm | -130dBm | -133dBm |
| **500kHz** | -116dBm | -119dBm | -125dBm | -125dBm | -128dBm | -130dBm |

In summary, a high SF offers a longer reach but at a lower data rate and increased transmission time; the inverse is true for a lower SF. Consequently, setting the SF for the

intended purpose is of great importance. SF is explored further in the design and development of the proposed localisation technique.

### *2.3.3 LoRa Frame Format*

A LoRa frame begins with a preamble that occupies the entire bandwidth and encodes a synchronisation word used to differentiate between networks that use the same frequency bands. An optional header (transmitted at 4/8 code rate) follows the preamble indicating the size of the payload, the code rate, and whether there is a 16-bit cyclic redundancy check (CRC) at the end of the frame core to enabling the receiver to check packets with correct headers. The payload size is stored using one byte, which limits the size of the payload to 255 bytes. The header is not required where payload length, coding rate, and CRC presence are already known. The structure of a LoRa frame is presented in Figure 2.8.


**Figure 2.8: Structure of LoRa frame.**

### *2.3.4 LoRaWAN Classes*

There are three classes of LoRaWAN devices: Class A, Class B and Class C, based on their load on the network [47]. All devices must have Class A functionality as a minimum to be considered LoRa-certified devices.

- **Class A, bi-directional:** devices with the lowest power consumption as they can be inactive for a long periods of time to conserve battery power. Given the only process

for the Gateway to communicate downlink with the end-device is to wait for an uplink transmission and then respond, they are less flexible on downlink transmissions.

- **Class B, bi-directional with scheduled receive slots:** in addition to the functionality of Class A, Class B devices have dedicated time slots for receiving downlink messages, and also periodically receive beacon messages in order to synchronise clocks.
- **Class C, bi-directional with maximal receive slots:** devices with the highest power consumption, given their open receiver windows enabling continuous reception [48].

Direct device-to-device communications is not possible; data can only be transmitted device-to-serve or server-to-device. Any device-to-device communication must be routed via the server.

## 2.4 Localisation of IoT Nodes

As the number of connected entities continues to increase, geo-localisation of nodes is becoming increasingly important for many IoT applications. Accurate localisation of nodes can be prohibitively expensive especially as their number proliferates. Thus, there is continued interest in developing cost-effective engineered solutions for accurate localisation within extensive IoT implementations.

### *2.4.1 Satellite-Based Location*

A natural solution to localisation is to equip each device with a Global Navigation Satellite System (GNSS) – “Global Positioning System (GPS) [49] or "Global Navigation Satellite

System (GLONASS)" [50] - capability. The addition, for example, of GPS interfaces to nodes harness GNSS systems to provide high localisation accuracy but at the expense of complexity, significant power consumption and cost. Another drawback of satellite-based systems is the inability to function indoors (offices, homes, factories and malls) as the path from the satellite to the node is impaired by thick concrete and metal structures. The more visible the path between satellite and node, the more accurate the localisation.

GPS is a widely used satellite constellation comprising 32 satellites [51] providing location and time information [52]; at least four satellites are visible simultaneously from any location across the globe. GPS uses signal timing in similar fashion to Time-of-Arrival (ToA) to determine the distance from the satellites, forming the basis for the estimation of location of the receiving node. GPS was introduced in 1973 by the Ministry of Defence of the United States Government to aid in military applications and subsequently has found extensive applications in civilian applications. In the early years of deployment, the military sector in particular, derived significant advantage through locating threats without being physically present at the location. It was also utilised in the monitoring of aggression and useful in informing on strategies to defeat enemies. The system attracted the interest of many countries and stimulated the launch of other satellite networks. Over the years, the scope of uses of GPS has grown increasingly owing to the advancement in technologies that have enabled the integration of receivers into devices such as mobile phones [53]. Other essential applications that mined GPS included time transfer for synchronisation and the timing control of traffic lights.

Although GPS consistently achieves high levels of positioning accuracy [54] through precise time synchronisation, its implementation requires a relatively high performance platform than routinely available in IoT systems; in the latter, nodes are usually equipped with very low-computing power that only perform basic operations. The power consumption requirements are onerous in this respect, multi-path can weaken GPS signals strength and obviates its use in indoor environments. Consequently, the operation of GPS for densely populated areas [55] is limited.

GPS has been considered for IoT node positioning applications. However, implementations have significant power requirements and are relatively expensive. Consequently, GPS-based solutions are generally not suitable for long-range, low-power IoT systems as the challenge is exacerbated given the available power on the sensor node is limited, compromising GPS accuracy and usability. [56] report that a device equipped with GPS and Global System for Mobile Communication (GSM) functionality consumed 30-40 times the power of a device equipped with LoRa only. In addition, as the GPS and GSM functions are separate, making miniaturisation difficult whilst also increasing costs. All these factors make GPS unsuitable for node localisation for the application considered in the research.

Assisted GPS (Figure 2.9) - as the name suggests - is a modification of the GPS to supplement the already established functionalities of GPS by facilitating faster and more efficient signal interpretation. Assisted GPS (AGPS) harnesses network resources to locate and use the satellites in poor signal conditions [57]. AGPS was developed to manage the problem of call time to first fix caused by GPS systems on activation. GPS systems require time to acquire signals from the satellite, navigate data, and perform localisation; AGPS consists of partial

GPS receivers, AGPS servers, a wireless network of base stations and a mobile switching center (MSC). The location of the handset is obtained from the MSC identifying the cell location as well as the sector of the handset through directional antennas at the base stations. Since AGPS servers track and monitor the satellites, a prediction can be drawn on whether a signal to the cellular devices is ongoing, facilitating the acquisition of signals when the GPS is activated, thereby reducing the time to acquire location significantly. Thus AGPS provides faster transmission of data and improves GPS satellite coverage. Currently the method is implemented mainly in environments with many obstructions such as high mountains and excessively deep valleys [58].

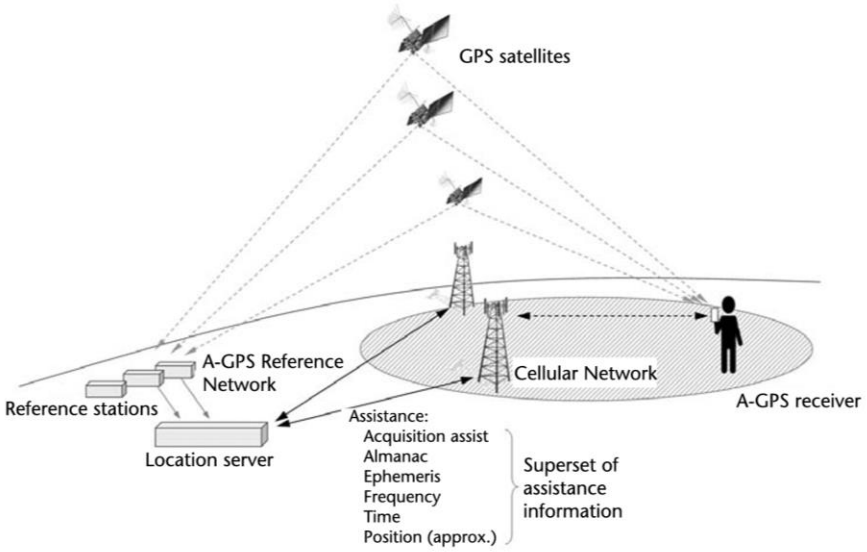



**Figure 2.9: Assisted GPS [57].**

IoT systems comprising nodes equipped with GPS to communicate and transmit information through an Internet connection have been implemented. Standalone towers transmit information from GPS satellites to areas deemed to be unreachable, facilitating faster

propagation of data, overcoming errors owing to multi-path [59]. The performance of the solution is enhanced by AGPS.

Although GPS and AGPS are the most common used positioning systems, they are unsuitable for certain environments because the accuracy is dependent on many factors such as the position of the satellite during data recording, surrounding buildings, valleys, and trees as well as the weather.

The main challenge for large scale IoT deployments in general, is power consumption and cost. The utilisation of a positioning technique derived from the network rather than a GNSS-based technique, yields a more cost effective solution. Furthermore, many networks are designed for indoor use, extending the range of applications.

### *2.4.2 Network-Based Localisation Methods*

Network-based location estimation relies on measurable Location Dependent Parameters (LDPs) which are then used to estimate the distance between transmitter and receiver nodes. LDPs that are commonly used are Time of Arrival (ToA), Time Difference of Arrival (TDoA), Angle of Arrival (AoA) and Received Signal Strength (RSS), forming the foundation for a number of frameworks to calculate the distance between nodes. Irrespective of the parameter considered, the model is configured to have certain nodes of known location known as “beacons”/“anchor nodes” or more commonly Gateways and other deployed nodes (the “end-devices”), the locations of which are to be determined.

*2.3.2.1 Time of Arrival (ToA)*

Time of Arrival (ToA) represents an absolute time taken for signals to arrive from transmitters to receivers and has been used as the basis to determine the location of a node [60, 61]. The distance to the anchor and any node is derived from the time using the known speed of light $c = 2.98 \times 10^8 m/s, d = c\tau$, where $\tau$ is the time of flight. Under multiple anchor nodes scenarios, the position of any node can be determined through trilateration. If the signal was transmitted at time $t_M$, the time to reception at anchor node $i$ would be $t_i$ as per Equation 2.3, where $d_i$ is the distance between the sensor node and anchor node. $i$, $x_i$, $y_i$ are the two-dimensional coordinates for anchor node $i$, and $x$, $y$ are the two-dimensional coordinates for a node;

$$t_i = \tau_i + t_M \qquad \textbf{Equation 2.3}$$

$$t_i - t_M = \frac{d_i}{c} = \frac{\sqrt{(x_i - x)^2 + (y_i - y)^2}}{c}$$

In the case of one anchor node only, the position of a node cannot be determined; with two anchor nodes, two nonlinear equations, Equation 2.4 and Equation 2.5 can be established:

$$t_i - t_M = \frac{\sqrt{(x_i - x)^2 + (y_i - y)^2}}{c} \qquad \textbf{Equation 2.4}$$

$$t_j - t_M = \frac{\sqrt{(x_j - x)^2 + (y_j - y)^2}}{c} \qquad \textbf{Equation 2.5}$$

In most cases, two solutions are evaluated i.e. locations, as depicted in Figure 2.10(a) (or one if the node is equidistant from the two base stations). For three anchor nodes, one solution (location) is obtained, illustrated in Figure 2.10(b).

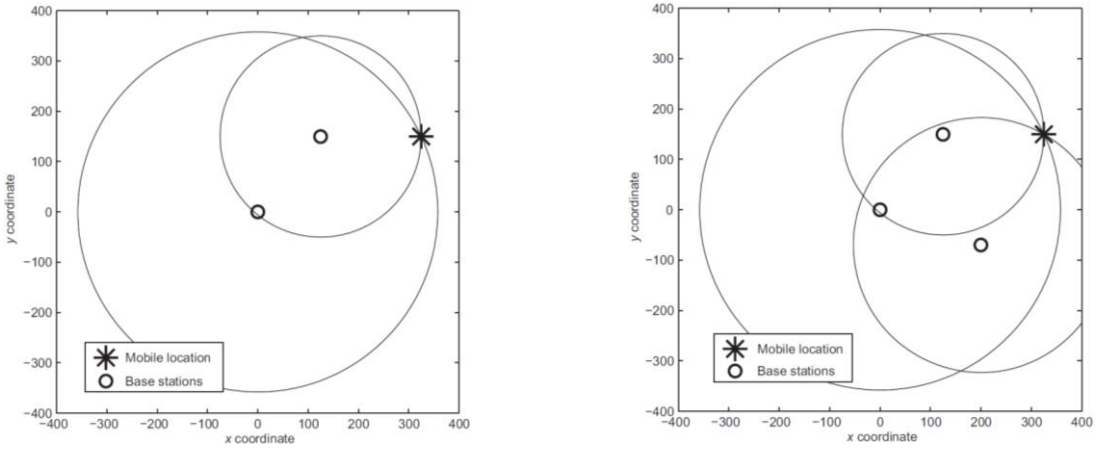



**Figure 2.10: Time of Arrival (ToA) position solutions [60].**

However, it must be stressed that ToA measurements require the anchor and receiving nodes to be synchronised [7]; in order to achieve accurate distance measurements, synchronisation needs to be at the nanosecond scale. For low cost, low power consumption networks such as LPWAN where nodes are idle most of the time, maintaining synchronisation is a significant overhead and is subject to "clock drift".

*2.3.2.2 Time Difference of Arrival (TDoA)*

Time Difference of Arrival (TDoA) [56] [62] is also a time-based localisation method introduced to circumvent the need for synchronisation of both Gateways and node clocks [60, 61]. TDoA is a measurement of the difference in the signal arrival times at two Gateways.

If the two Gateways have synchronised clocks, then multi-lateration can be used to locate the node.

If the signal was transmitted at time $t_M$, the time to reception can be related to the distance between the sensor node and anchor node $i$ as in Equation 2.6;

$$t_i = \tau_i + t_M \qquad \textbf{Equation 2.6}$$

$$= \frac{d_i}{c} + t_M = \frac{\sqrt{(x_i - x)^2 + (y_i - y)^2}}{c} + t_M$$

Of the three unknowns$x$, $y_i$, and $t_M$, $t_M$ can be eliminated by taking the difference between two arrival time measurements at two Gateways as in Equation 2.7 where $i \neq j$;

$$t_i - t_j = \tau_i + t_M - (\tau_j - t_M) \qquad \textbf{Equation 2.7}$$

$$= \tau_i - \tau_j$$

$$= \frac{d_i}{c} - \frac{d_j}{c}$$

$$= \frac{\sqrt{(x_i - x)^2 + (y_i - y)^2} - \sqrt{(x_j - x)^2 + (y_j - y)^2}}{c}$$

Rather than producing a circle of possible locations as in the ToA solution, the TDoA method produces a hyperbola of possible locations.

Similar to the ToA position calculations, additional equations are required viz. Equation 2.8 and Equation 2.9, which necessitate a third anchor node;

$$t_i - t_j = \frac{\sqrt{(x_i - x)^2 + (y_i - y)^2} - \sqrt{(x_j - x)^2 + (y_j - y)^2}}{c}$$ **Equation 2.8**

$$t_i - t_k = \frac{\sqrt{(x_i - x)^2 + (y_i - y)^2} - \sqrt{(x_k - x)^2 + (y_k - y)^2}}{c}$$ **Equation 2.9**

Two nonlinear equations with two unknowns may yield one or two solutions including the redundant hyperbola.

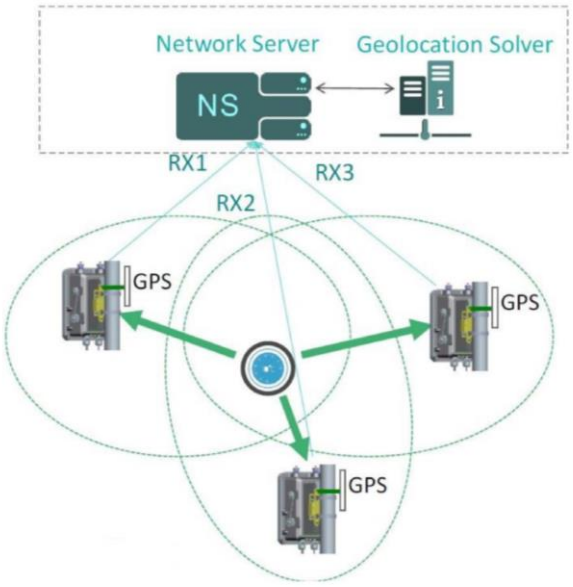



**Figure 2.11: Time Difference of Arrival (TDoA) solutions [63].**

Therefore, in order to ensure a single solution, three (non-redundant) hyperbolas from three anchor nodes are required as in Figure 2.11. ToA and TDoA methods require nanosecond timestamping accuracy for accurate localisation [64] [63].

*2.3.2.3 Angle of Arrival (AoA)*

Angle of Arrival (AoA) can be used as a means for determining the direction of propagation of a radio-frequency signal incident on an antenna array. The direction of arrival is determined by measuring the time difference of arrival (TDoA) through the difference in received phase at each individual element in the antenna array; in effect beamforming in reverse.

In beamforming, the signal from each element is weighted to "steer" the gain of the antenna array. In AoA, the delay of arrival at each element is measured and used to calculate the angle. Geo-location using AoA requires a minimum of two receivers. According to [65] , “AoA uses an antenna containing a multiple element array in which the exact location of each AoA element is known precisely.” The elements can receive separate signals that have different strengths. After measuring the strength of the signals, time of arrival, and different phases of every element, the path of a line of sight is calculated [66]. Another receiver with a similar antenna configuration is placed at different positions, and the procedure repeated such that the crossover of the two lines of sight indicates the location of the transmitter. Therefore, AoA uses triangulation founded on vector ranging. AoA-based solutions require large and complex antenna arrays, which also require complex periodic calibration [67] and are costly

One common application of AoA is in the geo-location of cell phones, most impactful in the reporting of the location of an emergency call or to provide location services to the user or a remote manager of the phone. Multiple receivers at a base station can measure the AoA of the phone to determine the phone's geodesic location. AoA has also been used to discover

the location of pirate radio stations or military radio transmitter [68]. In submarine acoustics, AoA has been used to localize objects with active or passive ranging.

#### *2.3.2.4 Received Signal Strength Indicator (RSSI)*

An alternative method of estimating location is by use of a path loss model that predicts the signal attenuation over distance. The Receive Signal Strength Indicator (RSSI) measures the signal power level at the receiver [69]. The measure varies with the transmitter-receiver distance and is highly dependent on the free-space propagation environment. The attenuation in RSSI is proportional to the inverse of the squared distance [70]; thus the distance to the sender can be estimated through the relationship of the strength of a signal against the model.

The estimated distance depends on the chosen path loss model, an approximation subject to interference, obstructions, reflections, absorption and multi-path fading. For example, a receiver obstructed by a wall will experience a significant reduction in signal strength and the distance calculated by the model will be much greater than is the actual case. Due to the significant complexities in accounting for all the factors impacting signal strength, empirically derived models are most often used. Such models are explicitly created for a certain type of environment, thus making a generalised model for all environments irrelevant. The derivation of models in order to capture the propagation characteristics of any particular environment is challenging.

A number of published path loss models are available for the determination of node location based on RSSI. The model is established by characterising the propagation of waves,

capturing the behaviour within a model and utilising a pattern matching method of the measured RSSI within a signal (radio) map.

The RSSI proximity method [47] can also be used to determine location depending on proximity of Gateways and received signal sensitivity; evaluation of the decrease in signal intensity indicates the location of a LoRa node. They note that the location can be estimated more accurately through the measurement of the signal strength.

RSSI based localisation methods are the most economical techniques for localisation since any additional infrastructure is not required and every radio chipset is equipped with a RSSI capability [71] [22] [16]. RSSI methods may lack the high accuracy of angle- and/or time-based techniques but offer the low cost and power consumption required by extensive IoT implementations. The advantageous characteristics motivates an increased interest in investigating RSSI based localisation methods with improved accuracy. The Thesis focuses on the development of a RSSI based LoRa localisation using fingerprinting and machine learning techniques for enhanced accuracy.

### *2.4.3 Related Work*

Table 2.2 summarises reported developments using Received Signal Strength Indicator (RSSI) based fingerprinting for the estimation of node location in LoRaWAN and SigFox settings. [22] details the development of fingerprint localisation from SigFox and LoRaWAN datasets acquired in large, outdoor environments ($52km^2$); the fingerprint was enhanced through K-Nearest Neighbour (KNN) methods. KNNs were also used to achieve a mean

distance error of Sigfox at 689 m and 398 m (LoRaWAN) form 84 and 68 base stations respectively , as part of Antwerp's 'City of Things Urban Environment' [72].

[71], which used the dataset of [22], provided analyses that proved the use of SVR for fingerprinting with LoRaWAN in urban deployments; the median accuracy of the SVR-enhanced estimation localisation errors was 250m. Interestingly, the location of the (SigFox or LoRaWAN) base stations was not provided nor the SF (for LoRaWAN results) used in the development. These 68 base stations were dispersed across city at unknown location with no detail on their relative positions.

[16] focused on the development of an outdoor parking positioning system for a restricted coverage area (340m x 340m) utilising 4 LoRaWAN base stations transmitting at SF7. Maximum Likelihood analysis achieved a mean distance error at 24m. Gaussian Process Regression (GPR)-based fingerprinting model for localisation [73] achieved a mean distance error of 25m in a campus outdoor area (150m x 250m) utilising 10 LoRaWAN base stations transmitting at SF12. These two studies focused solutions for relatively modest outdoor coverage areas.

The research reported in the Thesis extends the state-of-the-art by considering more extensive coverage areas (in the order of kms), harnessing the spreading factor in isolation or in combination - to optimise the estimation of location.

**Table 2.2: Related work in location fingerprinting.**

| Reference | Localisation Model/technique | Test environment | Technology | No. of Gateways | Spreading Factor | Estimation Error |
|---|---|---|---|---|---|---|
| **(Aernouts, et al., 2018)** [22] | Fingerprinting (KNN) | Outdoor (52 $km^2$) | SigFox | 84 | --- | Mean (689 m) |
| **(Aernouts, et al., 2018)** [22] | Fingerprinting (KNN) | Outdoor (52 $km^2$) | LoRaWAN | 68 | SF= 7 to 12 | Mean (398 m) |
| **(Choi, et al., 2018)** [16] | Fingerprinting (Maximum Likelihood) | Outdoor Parking (340mx340m) | LoRaWAN | 4 | SF=7 | Mean (24 m) |
| **(Zhe, et al., 2019)** [73] | Fingerprinting (Gaussian process) | Outdoor (150mx250m) | LoRaWAN | 10 | SF=12 | Mean (25 m) |
| **(Lemic, et al., 2019)** [71] | Fingerprinting (SVR) | Outdoor (52 $km^2$) | LoRaWAN | 68 | SF= 7 to 12 | Median (250 m) |

## 2.5 Summary

The background and review of technologies for IoT-based node location applications are discussed. A case has been made for the use of the LoRaWAN, developed for long-range, low power applications, as the most suitable technology for the proposed localisation application.

A general review of localisation techniques applicable for use within LoRa deployments is presented. A comparison of characteristics for three commonly used localisation techniques;

GPS, AGPS, ToA, TDoA and AoA are outlined; in the context of the focus of the research application, most are considered to be prohibitively complex and costly.

Thus the localisation approach adopted is the use of the routinely available RSSI to infer node location, detailed in subsequent Chapters.

# CHAPTER 3 RSSI MAPPING FOR PATH-LOSS CHARACTERISATION OF LORA

## 3.1 Introduction

The Chapter presents the methodology in executing a series of measurements for data acquisition in order to characterise the radio propagation of LoRa in clear and sandstorm environments across a suburban area of Jazan City, Saudi Arabia. The aim is to capture the transmission performance of a network of LoRa nodes, the basis for the determination of the feasibility and performance of RSSI-based estimation of node location.

The propagation of electro-magnetic (EM) waves is severely affected by objects in the path of the signal inducing increased levels of attenuation. There is a growing interest in the effect of dust particles on the propagation of radio/microwave signals brought about by the increasing number of terrestrial and satellite links for medium to long-range connectivity in regions dominated by dust and/or sandstorm [74]. In sandstorm conditions winds agitate significant quantities of sand particles and consequently the path visibility between receiver and transmitter is reduced at speeds between 10km/hr to 40 km/hr. Characterisation of these effects require knowledge of the properties of the scattering particles and climate conditions of the application environment.

The Kingdom of Saudi Arabia (KSA) occupies an extensive area of land and is regarded as a country subject to a severe climate. Wireless communication networks installed in such environments suffer from attenuation owing to scattering and absorption [75], resulting in significant path loss as more packets strengths fall below the receiver sensitivity.

In order to determine the path loss characteristics within KSA environments for medium to long-range connectivity applications, a series of systematic measurements was undertaken for two weather conditions; clear condition and sandstorm. LoRa nodes and Gateways were deployed across the evaluation site to gather real time measurements. The measured RSSI then forms the input parameter that models the effects of a sandstorm on signal propagation, the foundation for the proposed solution for node location estimation.

The Chapter details the propagation model to be evaluated together with the detail of the experimental infrastructure with focus on the characteristic of the deployed LoRa nodes, the experimental configuration and the spectrum of data captured. Finally, a comparison of path-loss in clear and sandstorm conditions is presented to illustrate the impact of the latter on signal propagation.

## 3.2 Propagation Models

Radio propagation models - otherwise known as path loss models - are empirical mathematical formulations based on measurements taken within a specific scenario or environment and are representative of radio wave propagation as a function of frequency, distance from point of measurement and environmental factors. The models predict the reduction in power of a signal as it propagates through a medium (communication channel) subject to specific constraints. In other words, the models capture the level that the transmitted radio signal is affected by the environment, frequency of transmission, the distance between and the height of, transmitter and receiver. The path loss is as a consequence of reflections, scattering, diffraction and absorption, central in the prediction

of the strength of the signal at the receiver. Therefore, radio propagation models are fundamental in the design of wireless communication systems [76]; the propagation paths in IoT networks place limits on the extent of the implementation. Three main factors which affect radio propagation in are;

1. **Environment:** buildings, trees, dust particles, fog causes multipath propagation of radio waves, which contributes to background noise, degrading the power fo the signal at the receiver.
2. **Interference:** due to other sources generating electromagnetic waves within the area, and the concurrent signal transmissions by different nodes.
3. **Transceivers:** transceivers sensitivity set by internal noise processes.

Low power signals of wireless are also more vulnerable to multi-path distortion.

The radio propagation model must account for all factors that affect the quality of signal. However, this is not a trivial task. A number of path loss models that approximate different propagation scenarios have been reported; here, path loss is modeled using a two-ray ground reflection model.

### *3.2.1 Two-Ray Ground Reflection Model*

A two-ray ground reflection model treats path loss of a signal between transmitter and the receiver in Line-of-Sight (LoS) scenarios. Figure 3.1 shows a typical two-ray ground reflection model consisting of two components; a LoS and a multi-path component, the latter as a result of a ground reflection. The two-ray model is considered because it has been

proven to provide a better prediction at long distances as compared to other models [77]. The two-ray ground reflection path loss model is expressed as in Equation 3.1;

$$P_r(d) = \frac{P_t G_t G_r h_t^2 h_r^2}{d^4}$$

$h_t$ = height of transmitter

$h_r$ = height of receiver

**Equation 3.1**

$$PL\ (dB) = 10 \log \frac{P_t}{P_r}$$

$$PL\ (dB) = 10 \log \frac{P_t}{\frac{P_t G_t G_r h_t^2 h_r^2}{d^4}}$$

$$d = 10^{\frac{PL(dB) + 20 \log(h_t\, h_r) + 10 \log(G_t G_r)}{40}}$$

where d is the distance in meters between the receiver and transmitter and Pr(d) is the power received at a distance d. Equation 3.1 is used to accurately predict the power received at distance d between the transmitter and the receiver. The model is a derivative of the well-known free space Friis transmission model given in Equation 3.2:

$$P_r(d) = \frac{P_t G_t G_r \lambda^2}{(4\pi)^2 d^2 L}$$

$P_r(d)$ = received power

$P_t$ = transmitted power

$G_t$ = transmitter antenna gain

$G_r$ = receiver antenna gain

$d$ = antenna horizontal distance

$L$ =system loss factor

**Equation 3.2**

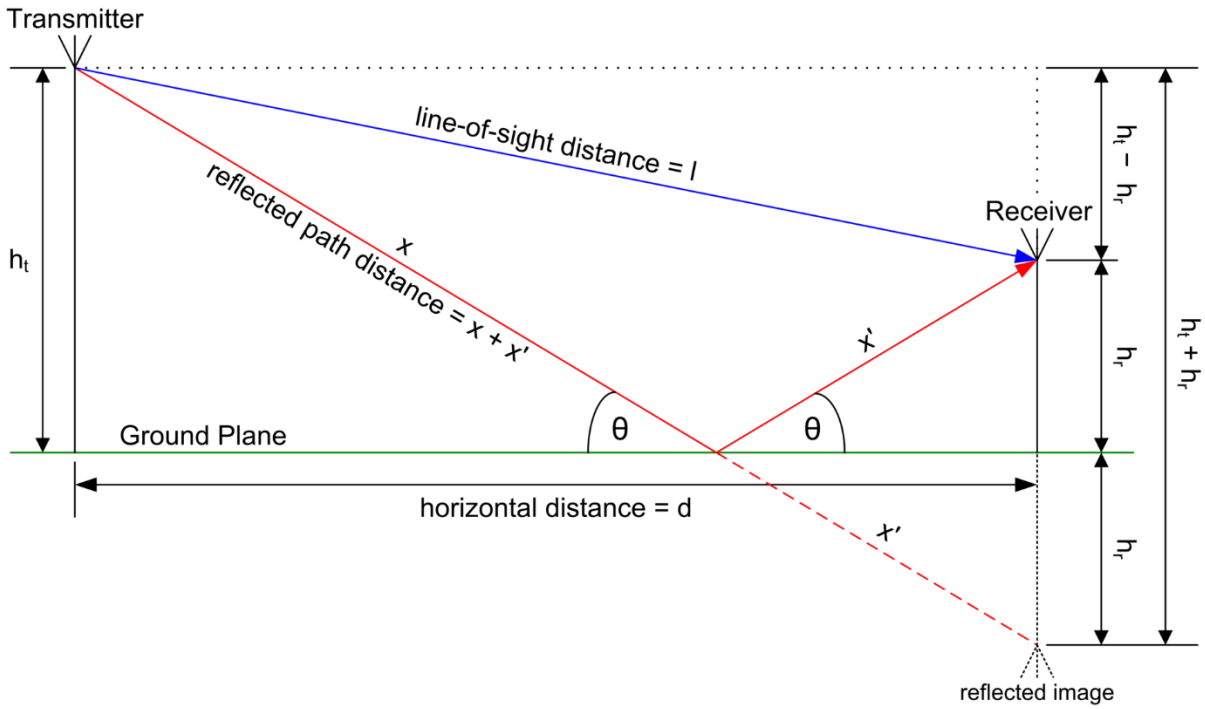



**Figure 3.1: The two-ray ground-reflection model.**

The two-ray model does not provide good predictions for short distances due to the oscillation caused by the constructive and destructive combination of the two rays. Instead, the free space model is still used when $d$ is small. Therefore, a breakpoint distance $d_b$ is calculated in this model. When $d < d_b$, Equation 3.2 is used. When $d > d_b$, Equation 3.1 is used. At the breakpoint distance, Equations 3.1 and 3.2 give the same result. So $d_b$ can be calculated as $d_b = {(4\pi h_t h_r)}/{\lambda}$ where $\lambda$ is the wavelength.

## 3.3 Data Acquisition

The series of measurements were conducted in a representative environment in Saudi Arabia - on the outskirts of Jazan near Jazan University – to characterise the radio propagation of LoRa in sandstorm conditions; Figure 3.2 shows the measurement area.

A receiver was placed on the roof of a car (approximately 2m above ground); the location of the transmitter node is taken with reference to the receiver (16.9610,42.5685). The transmitter node was placed in a car (approximately 1m above ground level) and moved to pre-defined positions of varying distances from the receiver. The extent of the area within which the measurements were recorded is characterised by direct paths; "Line-of-Sight (LOS)" paths were maintained throughout the measurements (actual visibility varied).

The Base Station (BS) comprised an *iC880ASPI LoRaWAN 868MHz* concentrator connected to a WiFi-enabled *Raspberry Pi 3 Model B SBC* platform (with 16GB micro SD card) and an *SMA* antenna 2 dBi (Figure 3.3.) The BS was housed in a *TEMBO ABS Enclosure* with a USB power connector and powered by a *Wopow* USB 5V 5000mAh battery. The Raspberry Pi is connected to the Internet via a *Zain Speed MIFI CAT6 4G* dongle with a laptop using *PuTTY* SSH Telnet client software monitoring the BS whilst receiving data and to subsequently download the data.

The transmitting node device was a *Sodaq One LoRaWAN* node with an *Anaren 868MHz* antenna 3 dBi powered via a USB port (in a car). The sensor was pre-programmed to transmit constant packets of data when connected to power; transmission was controlled crudely by connecting/disconnecting with the USB port. The transmitter power was set at 14dBm, initially at a Spreading Factor (SF) of 7 and the bandwidth 125 kHz.

In LoRa, the receiver sensitivity depends on the SF. Table 3.1 shows the receiver sensitivities for LoRaWAN, at each spreading factor for at a bandwidth of 125 kHz [46]. The total airtime to send the full packet at SF7 is 71.9 ms and the interval between two packets (at 1% duty

cycle) is 7.19 s. LoRa implements long-range connectivity by utilising a coding gain derived through spread spectrum modulation viz. a chirp spread spectrum (CSS) is used to modulate symbols over a fixed bandwidth. The SF governs the number of chips and range can be improved by increasing the SF. However, increasing the SF, reduces the data rate and the time on air is increased, requiring greater energy consumption. Consequently, the LoRa modulation permits a trade-off between range and energy consumption through selection of the SF.

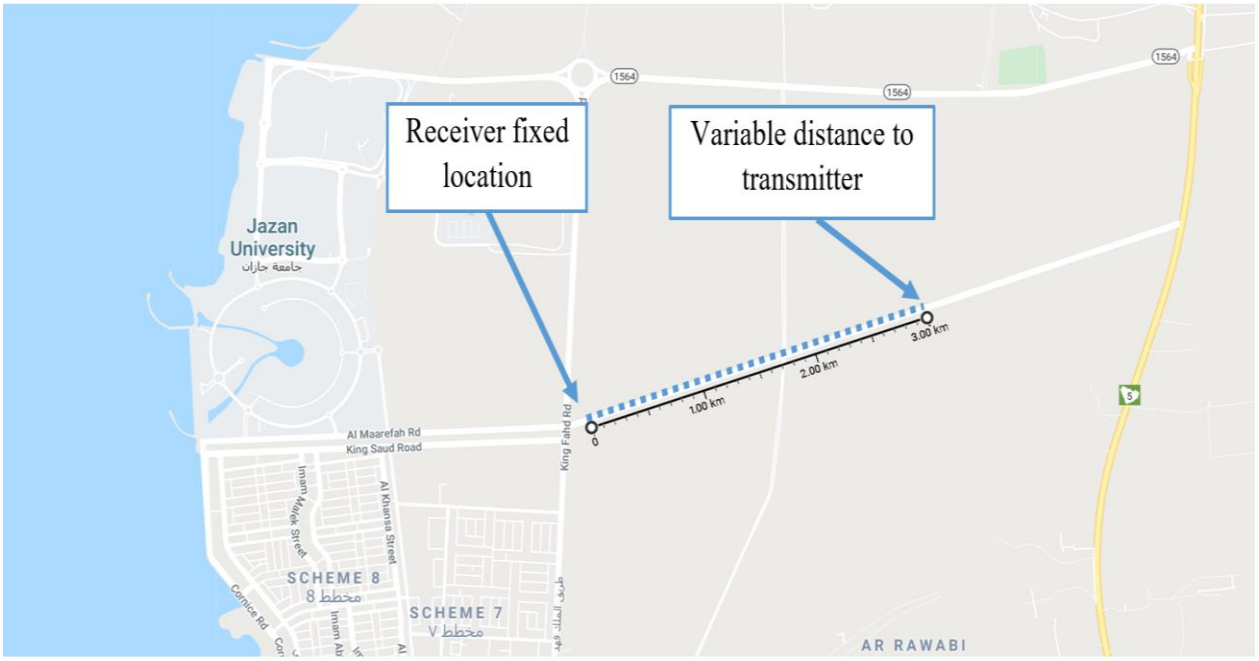



**Figure 3.2: The location where the series of measurements were recorded.**

**Table 3.1: Receiver sensitivity as a function of SF.**

| | SF7 | SF8 | SF9 | SF10 | SF11 | SF12 |
|---|---|---|---|---|---|---|
| **Sensitivity (dBm)** | -123 | -126 | -132 | -132 | -133 | -136 |
| | | | | | | |

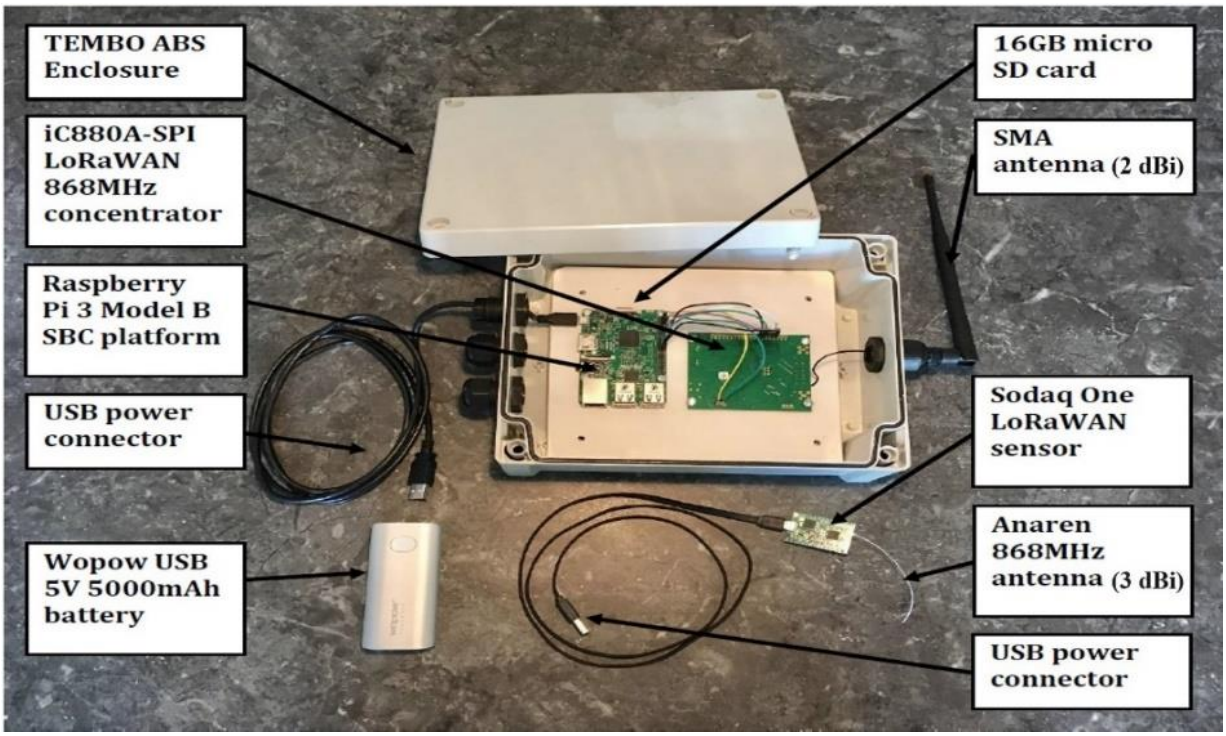



**Figure 3.3: LoRA Transmitter and Receiver modules.**

## 3.4 Data Acquisition Methodology

Measurements were carried out under two conditions; Figure 3.4 shows the two environmental conditions, clear and sandstorm. RSSI measurements were collected over two days under the two scenarios. For the clear condition, RSSIs were measured in the morning hours because during this period of the day, the wind speed are low.

The visibility between receiver and transmitter is compromised in sandstorm conditions. Thus RSSIs were obtained during afternoon hours, when the temperature was high, and the wind speed gave rise to sandstorms. Wind speed is the most critical environmental parameter that impacts signal propagation in this context. The wind stirs particles into the

air and as the strength of the wind increases, the density of the particles increases and the impact on the propagation of the radio signals becomes more significant.

**a – clear** **b – sandstorm**

**Figure 3.4: Environmental conditions (a) clear; (b) sandstorm conditions.**

At each location, the transmitter was connected to power until more than ten packets were received. If no packets were received, the transmitter remained powered for five minutes before the car-mounted transmitter was moved to the next location.

Table 3.2 and Table 3.3 show the number of packets successfully received at transmitter-receiver distances every 100m up to 3km under clear and sandstorm conditions. Evident is a significant decrease in the number of packets received with respect to distance from 1 km at clear and 1.5 km at sandstorm condition. Eventually no packet is received in the case of sandstorm, and as same as clear condition expect one packet at 2km.

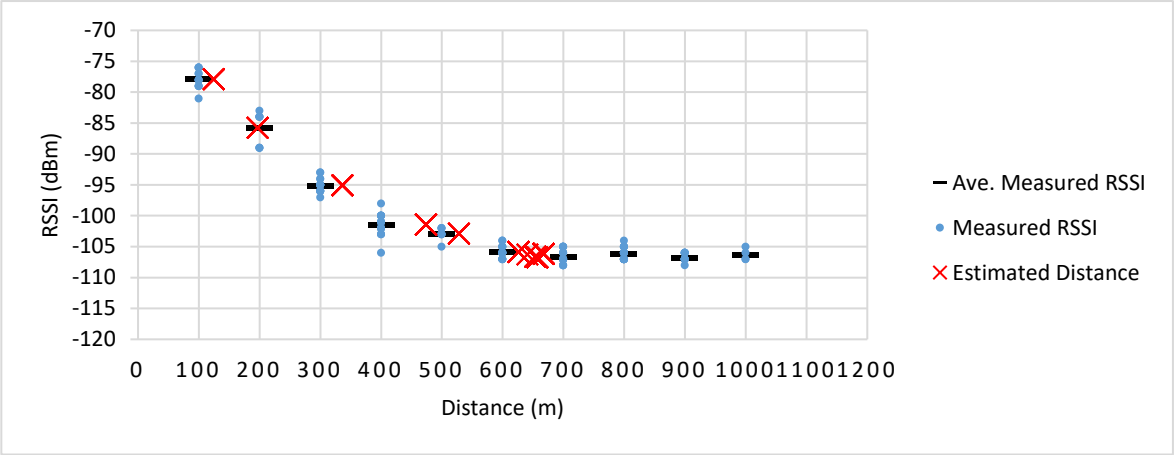
## 3.5 Analysis and Discussion

The Two Ray Ground Reflection Model defined in Equation 3.1 is used to estimate transmitter-receiver distance from the measured RSSIs in order to evaluate the path loss of the link. The mean of RSSI values shown in Table 3.2 and Table 3.3 are used as representative in the estimatation of the distances. Figure 3.5 (with raw data in Table 3.2 and Table 3.3) shows the variation in the actual and estimated distances with respect to measured RSSI. Evident, is the significantly higher signal attenuation in sandstorm conditions compared to clear condition. A radio signal in the presence of the sandstorm encounters an increase of the propagation path loss. The effect could be quite significant, depending on the location and weather condition at the time of measurement. The sudden drop after 500 m could be due to a sudden increase in the sandstorm given that measurements were taken at different times. Similar results have been obtained in [75]. Authors in [78] investigated the relationship between frequency and sand particles to determine the level of attenuation of radio signals in sandstorm conditions. They observed that the attenuation of microwave signals increases with increase in frequency for specific particle size. The attenuation varies from 0.0045 to 0.66dB/km at C-band and X-band frequencies when the humidity is equal to 0 %. While at humidity 60%, the attenuation varies from 0.023 to 3.93 dB/km. For Ku-band frequencies, the attenuation varies from 0.05 to 0.66 dB/km with a humidity equal to 0%, and from 0.13 to 9.78dB/km at humidity 60%. [78] established that attenuation due to sandstorm vary between 0.0045 to 9.78 dB/km at different frequencies (4 – 18 GHz) when humidity is 0 and 60% respectively. In [75], it has been found that attnuation between clear and sand storm condtions can be more than 10 dB/km for microwave radioations.

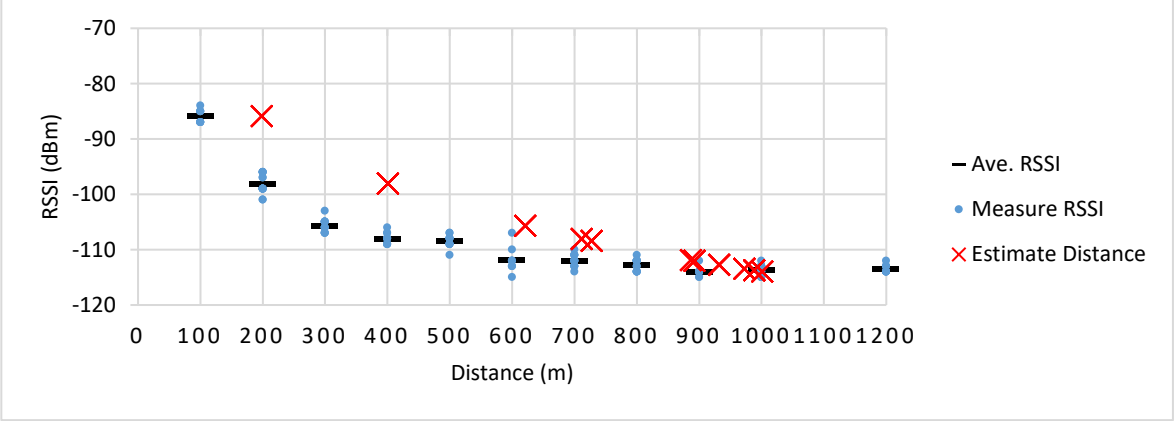
In [79],the authors investigated the expected attenuation at different frequency bands with respect to visibility and particle size in Saudi Arabia and Sudan. Their work shows that the attenuation varies from 0.22 dB/km to 0.0025 dB/km at S-band frequencies (2 – 4 GHz) for particle size of 50 micro-metre and visibility from 10 m to 500 m attenuation. At X-band (8 – 12 GHz), for same particle size and visibility range, the attenuation varies from 0.79 dB/km to 0.005 dB/km.

At Ku-band frequency (12 - 18 GHz) for particle size of 50 micro-metre and visibility from 10 m to 500 m attenuation varies from 3.8 dB/km to 0.01 dB/km. At frequencies in the K-band (18 – 26.5 GHz) the attenuation varies from 6.8 dB/km to 0.05 dB/km. At Ka-band (26.5 - 40 GHz) for particle size of 50 micro-metre and visibility from 10 m to 500 m attenuation varies from 13 dB/km to 0.2 dB/km. Also, at higher frequencies in the W-band (56 - 100 GHz), the attenuation varies from 47 dB/km to 2 dB/km for same particle size and visibility range.

However, in both scenarios, the signal attenuation appears to plateau within a certain range at longer distances. For the clear condition, the RSSI sensitivity is approximately -106dBm at 600m and greater; for sandstorm conditions, RSSI sensitivity is approximately -112dBm at 900m and greater. Consequently, a large proportion of estimated distances for clear and sandstorm conditions “plateau” at 600m to 700m and 900m to 1000m, respectively. It is also observed that in the sandstorm environment packets up to a distance of 2000m only can be recovered. Significantly higher attenuation of RSSI values are recorded in sandstorm conditions that in turn impact the estimation of distance.




**a – clear**




**b – sandstorm**

**Figure 3.5: Measured and estimated distances based on measured RSSIs using the two-ray propagation model.**

In sandstorm conditions, the path loss model (two-ray propagation) yields inaccurate estimates at shorter distances of up to 200m after which the estimated distances plateau and are significantly greater than actual distances. Even in clear condition, reasonable estimates are obtained up to 600m after which the estimated distances cluster around 650m, a significant underestimate of the actual distances.

**Table 3.2: Estimated distance error under clear conditions.**

| Actual Distance (m) | Mean Measure RSSI | Estimate Distance (m) | Est. Distance Error (m) |
|---|---|---|---|
| **100** | -77.85 | 124.93 | +24.93 |
| **200** | -85.75 | 196.91 | -3.09 |
| **300** | -95.06 | 336.50 | +36.50 |
| **400** | -101.41 | 485.00 | +85.00 |
| **500** | -102.89 | 528.13 | +28.13 |
| **600** | -105.85 | 626.27 | +26.27 |
| **700** | -106.60 | 653.91 | -46.09 |
| **800** | -106.16 | 637.55 | -162.45 |
| **900** | -106.69 | 657.39 | -242.61 |
| **1000** | -106.25 | 640.86 | -359.14 |
| **1200** | - | - | - |
| **1500** | - | - | - |
| **2000** | -112.00 | 892.31 | -1107.69 |
| **2500** | - | - | - |
| **3000** | - | - | - |

**Table 3.3: Estimated distance error under sandstorm conditions.**

| Actual Distance (m) | Mean Measure RSSI | Estimate Distance (m) | Est. Distance Error (m) |
|---|---|---|---|
| **100** | -85.90 | 198.62 | +98.62 |
| **200** | -98.10 | 400.88 | +200.88 |
| **300** | -105.71 | 621.40 | +321.40 |
| **400** | -108.08 | 711.93 | +311.93 |
| **500** | -108.46 | 727.87 | +227.87 |
| **600** | -111.89 | 886.62 | +286.62 |
| **700** | -112.00 | 892.31 | +192.31 |
| **800** | -112.76 | 932.47 | +132.47 |
| **900** | -114.00 | 1001.19 | +101.19 |
| **1000** | -113.78 | 988.46 | -11.54 |
| **1200** | -113.50 | 972.78 | -227.22 |
| **1500** | -113.00 | 945.18 | -554.82 |
| **2000** | - | - | - |
| **2500** | - | - | - |
| **3000** | - | - | - |

RSSI-based localisation methods require knowledge of the underlying radio environment such that a suitable propagation model that defines the relationship between RSS and distance between the transmitter and the receiver can be built. Detailed RF propagation modelling is non-trivial. However, there are standard propagation models such as log-distance that can be used to estimate source location [80].

In this work, it has been observed that the use of the two-ray propagation model with RSSI measurements to determine distances - and hence the basis for the estimation of the location of a node - under clear condition and sandstorm conditions only produces reasonable results up to 600m and inaccurate estimates at shorter distances of up to 200m respectively. This is grossly inadequate for the long-range sensor location applications considered in the Thesis. This motivates an investigation into an alternate method that does not depend on the two-ray propagation model.

## 3.6 Conclusions

A study that characterised the path loss of LoRa links in sandstorm conditions has been conducted. Two series of measurement were undertaken in Jazan City of Saudi Arabia to acquire representative coverage data in order to establish the effect of sandstorms on signal propagation. RSSI measurements were recorded at different transmitter-receiver distances in both clear condition and sandstorm conditions using commercially available LoRa devices.

As expected, a higher path loss is experienced (higher levels of packets lost) during sandstorms. Results indicate that that two-ray propagation model can only estimate distances up to 600m in clear condition and the model fails at shorter distances of less than 200m in sandstorm environments.

In conclusion, the use of the two-ray path loss model is not a viable approach for LoRa localisation for extensive IoT implementations, the target distance being 3km. The results motivate an investigation into the use of fingerprinting for node location, to be discussed in the next Chapter.

# CHAPTER 4 RSSI-BASED FINGERPRINTING FOR IOT NODE LOCALISATION

## 4.1 Introduction

The Chapter investigates the use of a Received Signal Strength Indicator (RSSI) based fingerprinting technique for the estimation of the location of nodes within LoRaWAN in a sand storm environment. Results from the Characterisation of the path loss detailed in Chapter 3 confirm that the Two-Ray Reflection Model yields meaningful estimates a distance of less than 100m due to multipath, shadowing and reflection and is thus not appropriate for more extensive application scenarios. Thus the use of fingerprinting in the goal of location estimation in extensive IoT networks is formulated. The results of experiment/field trial carried out in sand storm conditions in the city of Jazan in Saudi Arabia for this investigation are presented.

RSSI is used as location fingerprints for node localisation. Machine learning algorithms have been identified and employed to model the complex RSSI-location relationship and hence enhance the accuracy of estimating node location. Furthermore, an investigation into the impact of different spreading factors in the estimation of localisation is carried out. Finally, the results of a performance evaluation of the developed solution presented with a particular emphasis on accuracy.

## 4.2 Problem Statement

Several techniques used for sensor localisation are based on Time of Arrival (ToA) [7] Time Difference of Arrival (TDoA) [56], Angle of Arrival (AoA) [81] or RSSI [82]. These methods (ToA, TDoA, AoA or RSSI) are used to estimate the transmitter distance from the receiver and trilaterate or triangulate for location. ToA and TDoA are techniques that require stringent time synchronisation, a costly overhead on the implementation; both ToA and AoA techniques require Line-of-Sight (LoS) paths for acceptable accuracy. The radio propagation dynamics in the environment under study suffer from different degrees of multipath fading, shadowing and interference. The environment is best described as sub-urban and rural governed by the type of terrain, therefore, the requirement for LoS may be difficult to maintain in this application. RSSI uses a defined path-loss model to estimate location, however at the cost of location accuracy, demonstrated in Chapter 3. The limitations of these techniques motivates an investigation into the feasibility of the RSSI-based fingerprinting for long-range node localisation. Radio fingerprinting based location estimation techniques have been proven to be more reliable because the 'fingerprints' explicitly capture the dynamic of the environment for which the system is being designed. In the Chapter, an investigation into node localisation in sand storm environment based on RSSI-fingerprinting in conjunction with the nearest neighbour algorithm [26] and its variants are presented.

## 4.3 Experimental Procedure

A series of field measurements was executed using LoRaWAN devices deployed in Jazan City in Saudi Arabia to acquire the necessary data for the development and evaluation of the

proposed RSSI-based fingerprinting technique for sensor localisation in sand storm environments.

### *4.3.1 The Environment*

Jazan City is located along the Red Sea coast in the south west of Saudi Arabia. The region is categorised by sand dunes, mountains, as well as coastal areas resulting in significant climate diversity. The climate is affected by the tropical wind and varies owing the diversity of the surface and geographical characteristics of the region. The coastal plain (Jazan City) is temperate in winter, hot and humid and subject to frequent sand storms in the summer. The temperature rises from June to September, the mean temperature ranging between 25°C in January and 40°C in September, with a maximum of 46°C. The Relative Humidity (RH) increases from the eastern part of the plains to the west, ranging between 61% In July to 79% in December, the maximum reaching 99% with a minimum of 27%.

Monsoon winds last between June to September creating sand storms during the summer, rising to over 37 km/h. During strong sand storms, visibility is less than 100 m; so, any IoT location system will be most challenged in dust environments, the signal strength impacted most significantly due to increased levels of scattering owing to dust particles in the air. Most environments in Saudi Arabia are similar to the Jazan region.

Given these yearly geographical trends, measurements were carried out during the months of July and August, the most challenging season, characterised by dust, high temperature and humidity. Apart from the climatic factors, the propagation environment also features trees and buildings.

Table 4.1 represents the weather conditions during measurement in the month of July and August.

**Table 4.1: the weather conditions during measurement (July-August)**

| | **Humidity (% )** | **Temperature (˚C)** | **Wind speed (km/h)** |
|---|---|---|---|
| **Dust sky** | 60 - 85 | 35 - 45 | 5 - 10 |
| **Sandstorm sky** | 60 - 85 | 35 - 45 | 13 -27 |
| **Strong Sand storm** | 60 - 85 | 35 - 45 | 37 |

The propagation path between the test area and the gateways is characterized by buildings of different elevations (9 m - 30 m). Specifically, the path leading to gateway 2 comprises many buildings with higher elevation compare to the other gateways.

### *4.3.2 Data Acquisition*

The data acquisition system consisted of four LoRaWAN transceiver Gateways and transmitter accessing the Internet through laptops. The Gateways comprise *iC880ASPI LoRaWAN 868MHz* concentrators connected to a Wi-Fi enabled host (*Raspberry Pi 3 Model B SBC* platform with 16 GB micro SD card) via a *SMA* antenna and are housed in *TEMBO ABS Enclosures* with mains electrical power supply. The enclosures are designed to guarantee operation between -5°C to +55°C, meeting the requirements of the operational environmental conditions. Gateways are the data collectors of the architecture utilising 868 MHz channels for data transmission. Packets can be received from different nodes with different spreading factors, up to 8 channels in parallel. Gateways are also equipped with an

external control microprocessor and an RPi 3 unit is connected to the IMST concentrator via the SPI bus. The RPi 3 is Wi-Fi enabled and connected to 4G connectors in order to receive and transmit data to the server ('The Things Network' server [83]). Transmitter nodes are a Sodaq One v2 LoRaWAN device with an 868 MHz antenna connected to a GPS module (Ublox Eva 7M). The node consists of an RN2483 transceiver with 14 dBm transmission power and bandwidth of 125 KHz powered by an 800 mAh lithium battery. An annotated photograph of a transceiver Gateway and transmitter is as shown in Figure 4.1.

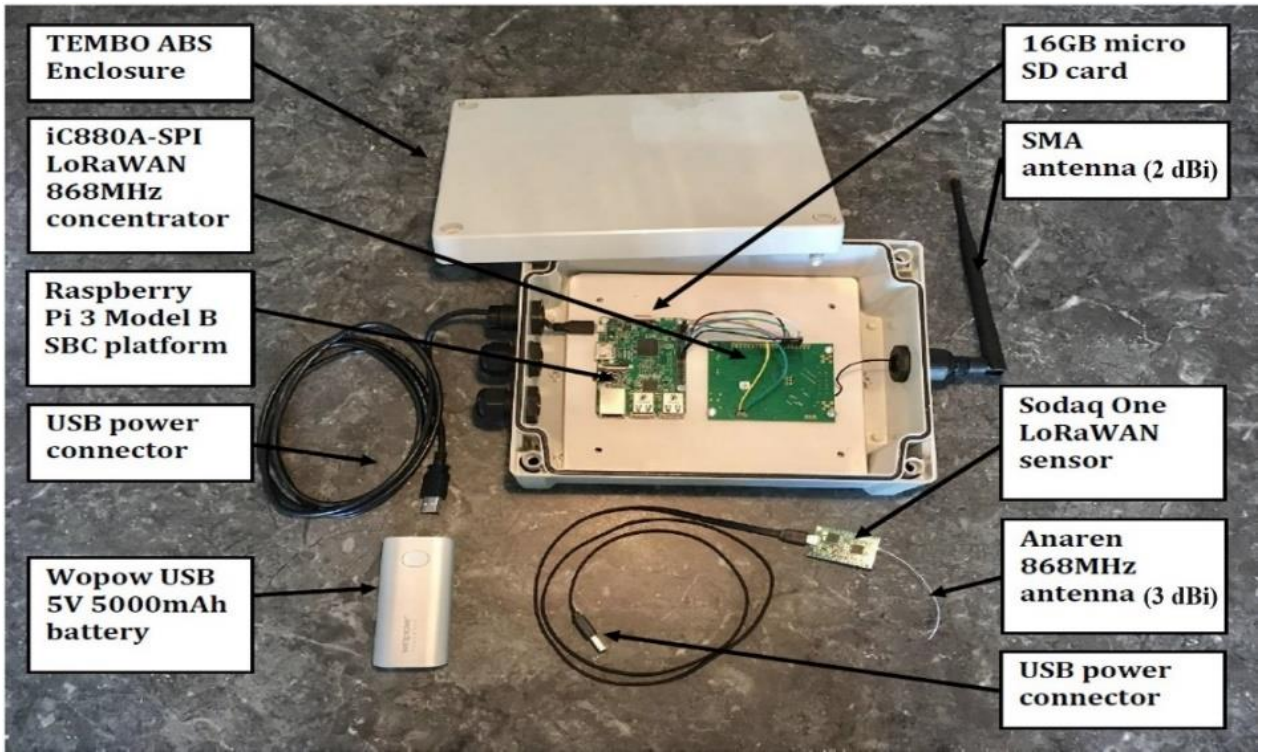



**Figure 4.1 LoRaWAN Receiver and transmitter.**

The Things Network (TTN), an open-source, de-centralised network designed to enable low power devices using long-range Gateways to exchange data with applications, was used to manage the received data from the Gateways. TTN permits a large variety of third-party

devices to be connected and provides a web interface to visualise data. However, communication with a third-party application is required if the objective is to process and analyse the data; to this end, an MQTT client forwarded the data to a third-party application. TTN uses MQTT to publish device activations and messages. However, a subscription developed in Python with appropriate device identifiers captured all packets from Gateways. The third-party application contained two principle components; Local MQTT Porker and PostgreSQL database. The main functions are to obtain data from the TTN; parse and insert the data into the specific Local MQTT Porker; then insert the data to the final destination the PostgreSQL database for processing. Local MQTT brokers organised the data flow and prevented packet collisions before storage in the database. The architecture of the deployed system interconnection is as shown in Figure 4.2.

Two laptops were utilised; the first as system server contained the PostgreSQL database where all the data fingerprints were stored. These data came from three MQTT local brokers (three RPi) connected through Wi-Fi. The first laptop also contains PuTTY software that issues commands and receives text responses over a TCP/IP secure socket (SSH)). PuTTY enables effective control of the system. The second laptop connected to the Internet (through a mobile phone) was used in the field to monitor node behaviours. PuTTY software was also installed on the second laptop for connection to the first laptop (server) to monitor and track transmitted packets and to ensure data is stored in the database. In addition, the three RPi 3 containing local brokers were checked periodically. Arduino board IDE software was used to program the node (Sodaq One v2).

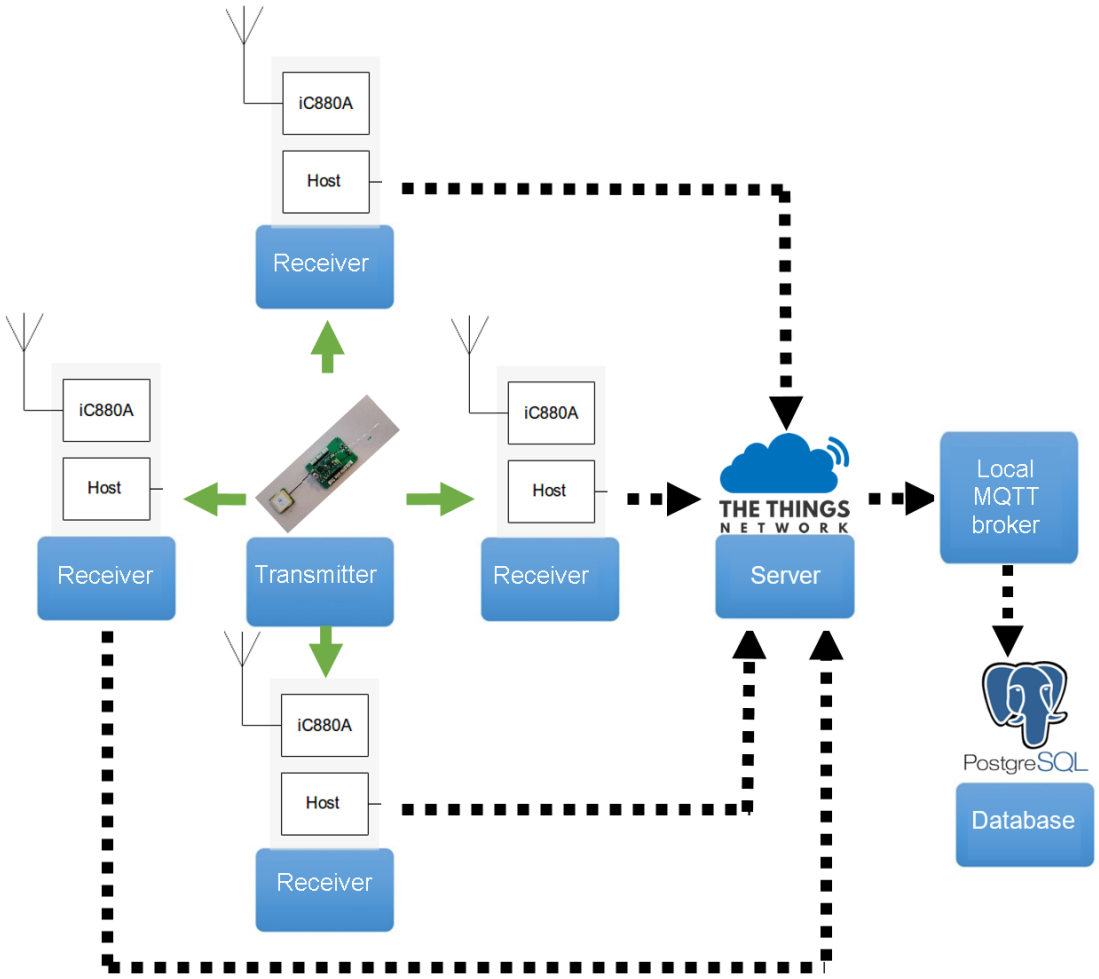



**Figure 4.2 Data acquisition system architecture.**

### *4.3.3 Field Experimental Set-Up*

The deployed configuration for the field trial is a non-uniform grid given the cluttered terrain (buildings, trees) of the environment. The map of the layout is as shown in Figure 4.3. The Gateways are placed on the outskirts of the City on four elevated sites shown in Figure 4.4 with their respective heights provided in Table 4.2.

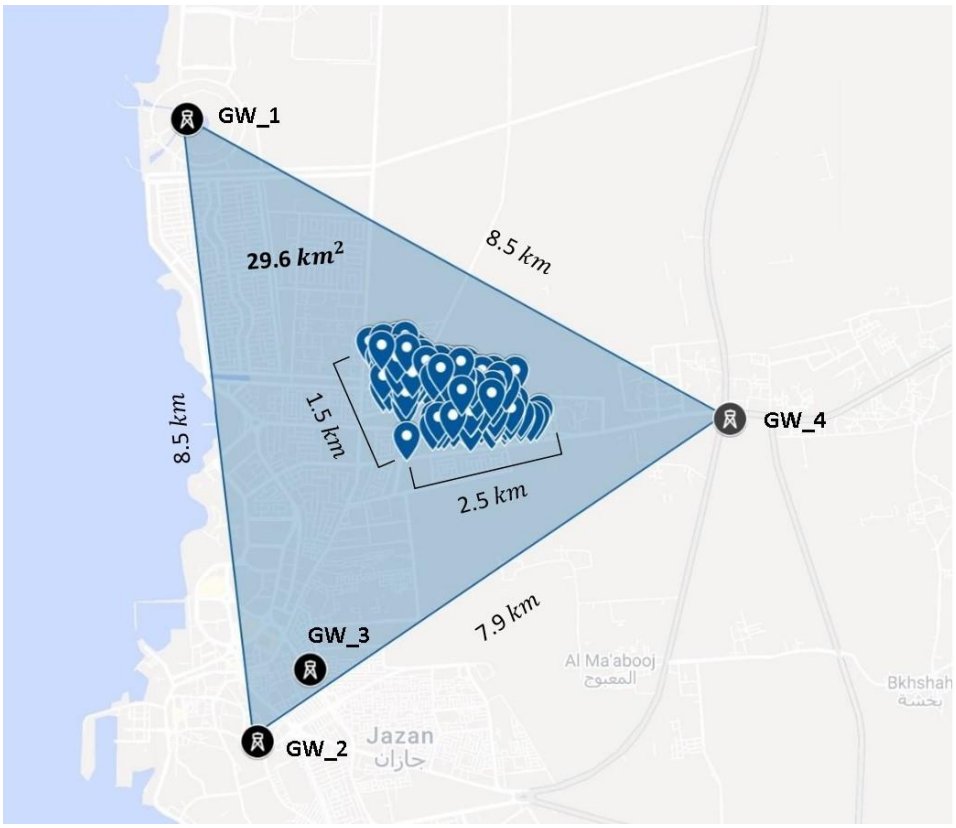



**Figure 4.3 Test field map.**

**GW_1** **GW_2** **GW_3** **GW_4**

**Figure 4.4: Gateway locations.**

**Table 4.2: Gateway locations.**

| Receiver | Location | Elevation (m) |
|---|---|---|
| **GW1** | Top of University tower | 100 |
| **GW2** | Communication tower (1) | 90 |
| **GW3** | Communication tower (2) | 70 |
| **GW4** | Top of Water tower | 40 |

Gateways are placed on elevated platforms to extend the range of the network that would otherwise be reduced due to the built environment and natural obstacles that influence the RSSI. The transmitter is fixed at one location when acquiring measurement data and then moved between measurement locations within the target area. Two series of measurements were carried out; the first set of measurements were taken from 40 locations using SF of 7; the second set were taken from all the 150 locations using SFs of 9, 10, 11 and 12. The distance between locations is approximately 100m. The distance between the closest point and Gateways is 4 km and the furthest is 7 km, the area of this experimental is $30km^2$. 20 RSSI packets are collected from each measurement point. Therefore, the two sets consist of a total of 800 and 3000 measurements respectively. The choice of SF=7 used for the first experiment was to test the suitability of the default SF in the LoRa device used for the study reported in this thesis. Higher SFs of 9, 10, 11 and 12 were chosen for the second experiment

with the expectation that they will provide the needed transmission and reception at long range.

### *4.3.4 Data Collection*

In the first set of measurements, the data collected utilises the LoRaWAN default Spreading Factor (SF) of 7. At each location, two transmitters were active simultaneously, broadcasting packets of data with the GPS location coordinates as a payload. Gateways that successfully acquired messages acknowledged receipt. The total airtime to transmit a packet at SF7 is 71.9 ms and the interval between two packets at 1% duty cycle is 7.19 s. The RSSI of the received packets is measured at each receiver and uploaded to TTN server along with the payload information. 20 packets were transmitted at each location at a transmission power of 14 dBm. (Table 3.1) shows the sensitivities for LoRaWAN, at each spreading factor for bandwidth 125 kHz.

The total number of packets acquired by each receiver for all 40 locations is shown in Figure 4.5; it is worth noting that just 6 packets were received by GW2. The significant loss of packets at GW2 may be due to the relatively long distance between the transmitter and receivers characterised by Non-Line-of-Sight (NLOS) paths. The Spreading Factor (7) also impacts reception; therefore, an investigation into the use of different SFs to improve reception was carried out.

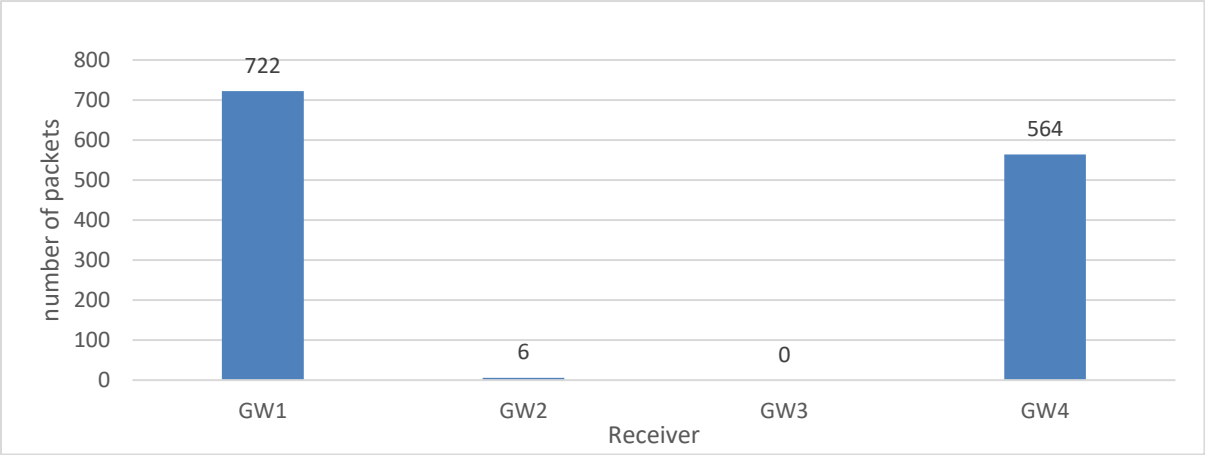



**Figure 4.5 Packets received by Gateway.**

### *4.3.5 Data collection as a function of Spreading Factor*

In order to understand further the context of the acquired data, consideration was given to the effects of Spreading Factor and the implications of The Things Network (TTN) usage policy.

The TTN Fair Access Policy limits the data each end-device can transmit, by allowing a mean of 30 seconds uplink time on-air, per day, per device [84]. A Spreading Factor (SF) can be set before each signal is transmitted, effectively spreading the signal over a wider spectrum range. As noted in the design, the basic principle of spread spectrum is that each bit of information is encoded as multiple chips [85]. As the SF increases, the time on-air, or symbol duration, for each packet also increase, as per the Equation 4.1. The effect of the TNN policy as a function of SF is shown in Table 4.3.

$$T_S = \frac{2^{SF}}{BW} \qquad \textbf{Equation 4.1}$$

**Table 4.3: The implication of LoRaWAN airtime policy as a function of spreading factor.**

| Spreading Factor | SF12 | SF11 | SF10 | SF9 |
|---|---|---|---|---|
| Packet duration (ms) | 1482.752 | 823.296 | 370.688 | 205.824 |
| Duty Cycle 1% (s) | 148.275 | 82.329 | 37.068 | 20.582 |
| Interval Time (mins) | 2.47125 | 1.37215 | 0.6178 | 0.343033 |
| Mean Number of Messages (/hour) | 20 | 36 | 80 | 145 |

For example, at the lowest data rate (SF12 and BW125 kHz), an 11-byte payload would need 1482.75ms of total airtime to transmit a full packet [86]. The 1% duty cycle limits the transmission of one packet every 148.275 seconds (or 2.47 mins); so the node needs to wait 2.47 mins before it may attempt to send another packet (in the same sub-band), and in tandem with the 30 seconds/day Fair Access Policy, only 20 packets can be sent per hour.

Packet size varies between 51 bytes for the slowest data rate (SF=12) and 222 bytes for fastest rates (SF=7). Therefore, it is beneficial to keep the application payload under 12 bytes and the interval between messages to be at least several minutes. Consequently, a conservative approach is not to transmit more than the smallest maximum payload size, which is 36 bytes; however, a loss of capacity results if a large amount of data has to be transmitted as well as lower throughput.

For the data collection, therefore, in order to use SF12, the payload was set to 11 bytes consisting of the GPS coordinates only, giving a time between packets of 2.47 minutes.

Therefore, in order to collect the volume of data required to train machine learning algorithms, the number of nodes/transmitters was increased.

The second series of measurements - conducted in a sand storm condition - acquired the RSSI of received packets at 150 different known locations using different SFs (9, 10, 11, and 12) referred henceforth as SF9, SF10, SF11 and SF12 respectively. As shown in Figure 4.3, the Gateway (black circles) were located at various points 4km to 7km around the area where the nodes were positioned (blue markers). As before, the Gateways were located on elevated platforms (Figure 4.4).

At each transmission location, 20 end-nodes were transmitting simultaneously at various spreading factors; three at SF9, three at SF10, six at SF11, and eight at SF12. The nodes transmitted packets containing GPS location coordinates as a payload. The RSSI of the received packets were recorded at each Gateway and uploaded to The Things Network server (along with the payload information; each was 11 bytes). 20 packets were transmitted at each location at each spreading factor value.

The total number of lost packets at each Gateway for each spreading factor for all locations is shown in Figure 4.6.; the blue, red, grey and yellow bars represent SF9, SF10, SF11 and SF12 respectively. A clear trend in the number of packets received as the SF value increases is evident; for each Gateway (GW1, GW2, GW3, GW4) the number of lost packets decreases as the SF increases. The differences are significant at GW2, GW3, and GW4. Although the trend is still evident at GW1, the difference is minimal for different SF values, potentially

owing to the fact that GW1 is located at the highest elevation (100 m) in the city providing relatively clear LOS paths with limited obstacles between GW1 and transmitter nodes.

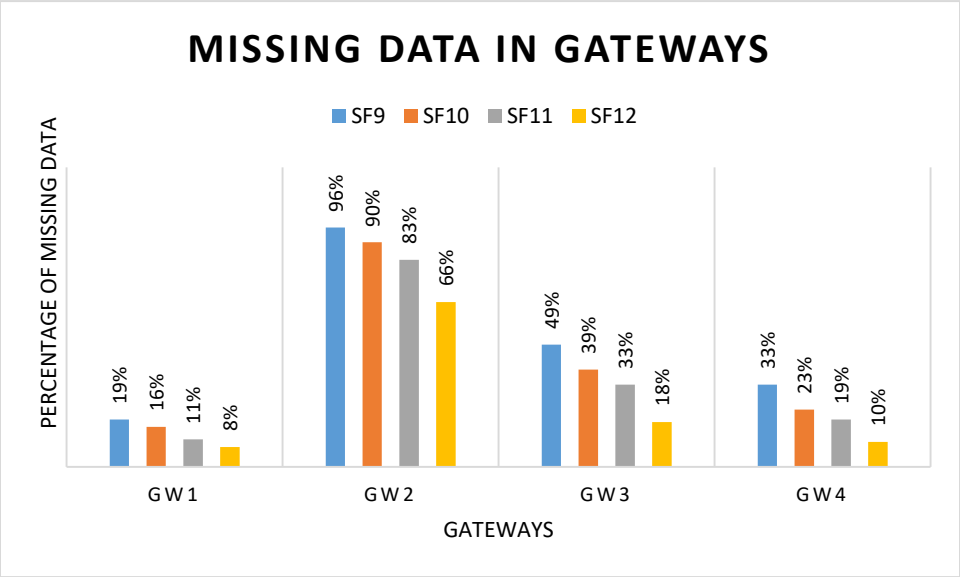



**Figure 4.6 Missing packets at each gateway**

### *4.3.6 Data Preparation*

The first stage in the process of preparing the data acquired for localisation is cleansing. These data comprise a range of information such as port, SNR, channel, frequency, which are not relevant for the purposes of the study; such entries are removed. The datasets are filtered such that only the necessary data viz. time, RSSI_Gateway1, RSSI_Gateway3, RSSI_Gateway3, RSSI_Gateway4, longitude and latitude are retained. RSSI values and their known locations (longitude and latitude) are central to the proposed node localisation techniques reliant on machine learning. The locations are core to partitioning the data into training and testing sets.

In the case of missing data, represented in Figure 4.6, the mean imputation method [87] is applied for filling such data since it has been proven to be more efficient than other methods such as Hot-Deck, Cold-Deck, Maximum-Likelihood, and no unit is sacrificed. In addition, if the observed data contain useful information for predicting missing values, an imputation procedure can exploit it and maintain integrity with high precision. Imputation provides a complete data set amenable to analysis by standard methods [88].

Imputation reconstitutes missing data as follows:

**Step 1**; Separation of each group of 20 packets by location.

**Step 2**; Examination of all data for each Gateway in isolation. Two issues are faced in respect of dealing with missing data:

1. In the case of loss of all the RSSI (lost packets) at a specific Gateway (referred to as Monotone) with the same location, missing RSSI values are replaced with the specified value, assigned as the high sensitivity value of -132 dBm.
2. In the case of a partial missing data at a specific Gateway (referred to as Non-Monotone) with the same location, the mean of observed RSSI values are calculated (not Null) of the $G_j$ for each location, where $G_j$ denotes the number of the Gateway i.e. these missing value are replaced with the mean value in $G_j$ for each location.

The resulting data contains 6 columns [RSSI_Gatewy1, RSSI_Gatewy2, RSSI_Gatewy3, RSSI_Gatewy4, Longitude, Latitude] and 150 rows where RSSI's are measured using 4 Gateways (receivers) from 150 different locations (longitude and latitude).

The study reported in this thesis was carried out using the open source scikit-learn machine learning library for the Python language. scikit-learn provide various tools for model fitting and prediction. Table 4.4presents the scikit-learn and python library with the algorithms used in this study.

**Table 4.4: Python Library and models**

| | Library | Model |
|---|---|---|
| **1** | python | Numpy, pandas, math, scipy, matplotlib.pyplot |
| **2** | sklearn.model_selection | KFold |
| **3** | sklearn.neighbors | KNeighborsRegressor |
| **4** | sklearn.svm | SVR, NuSVR |
| **5** | sklearn.gaussian_process | GaussianProcessRegressor |
| **6** | sklearn.gaussian_process.kernel | RBF, Matern, RationalQuadratic, ExpSineSquared, DotProduct, ConstantKernel |
| **7** | sklearn. tree | DecisionTreeRegressor |
| **8** | sklearn.ensemble | GradientBoostingRegressor, RandomForestRegressor |

The machine learning methodology, separate datasets are needed to train and validate the model. Here, the RSSI /location data were collected by four gateways from 150 locations, and then randomly divided into 80% for training and 20% for testing. The RSSI/location in the

training data were used as input/output feature for training the machine learning model. To validate the model the RSSI measurements from unknown location were used in the model to estimate the unknown locations. The accuracy of the models is measured by the Haversian distance metric between the estimated location and the true location of a node.

## 4.4 Location Fingerprinting

Location Fingerprinting (LF) is a technique that utilises any unique characteristic of radio signals that can be differentiated to infer the location of a node. LF exploits the relationship between radio signal transmission behaviour and known spatial locations to establish a model that can in turn be used to determine unknown locations of other signals. The commonly used signal parameter is absolute RSSI from multiple receivers. RSSI is known to have an inverse relationship with distance from transmitter to receiver [89].

In dynamic environments, such as the one under study, signal propagation is especially affected by multipath, limiting approaches to localisation that largely depend on LOS ineffective; conversely, multipath in the case of LF creates unique signatures for different locations. The development of LF can be crudely divided into two phases; an 'off-line' stage, also known as database creation and an 'on-line' stage known as location estimation [90] . A schematic diagram of fingerprinting localisation is as shown in Figure 4.7.

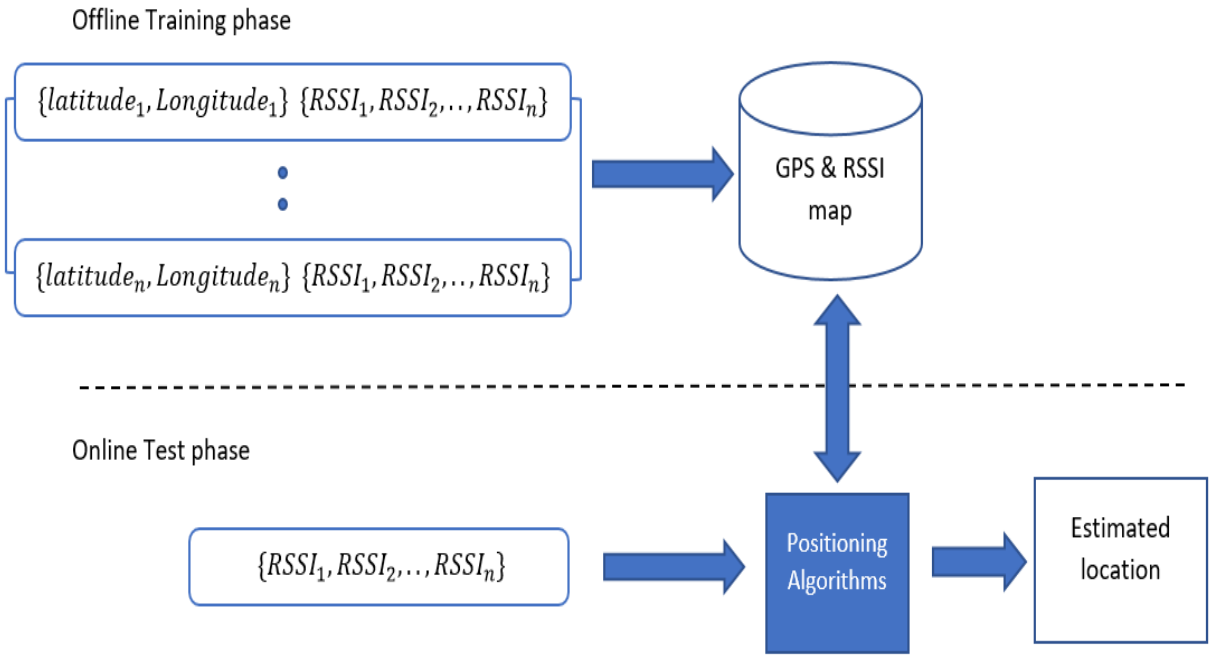



**Figure 4.7: Schematic of the Location Fingerprinting technique.**

First, a set of pre-defined grid points referred to as reference points are selected. During the off-line stage, a survey is carried out and multiple measurements are taken at each grid point throughout a time interval. The database of measurements recorded from all grid points allows the creation of a radio map for the area. In the location estimation stage, new measurements from other locations assumed to be unknown are taken and a location algorithm is used to match these measurements to the radio map entries to find similar fingerprints and the reference locations of those fingerprints are then used to estimate the locations of the new measurements. An important feature of location fingerprinting is that the locations of the receivers do not need to be known; only the received signal/packets for known transmitter node locations need to be known (in off-line stage).

Variations in the propagation environment is a huge challenge in the location fingerprinting. A periodic update strategy for the created database and refresh the radio map is implemented to cater for the changes that arise in the operating environment.

## 4.5 RSSI Location Fingerprint

RSSI is selected as the fingerprint parameter to implement localisation. Given that the radio propagation characterisation for the sand storm environment under investigation confirms that packets that contain RSSI information are subject to significant loss, the number of returned RSSI values is lower as is temporal variations in the RSSI values; this constitutes noise in the data. Therefore, in order to obtain a high-quality dataset, the RSSI information collected during the experiment is pre-processed to obtain a robust set of features to enable models for node localisation to be established.

Here, the mean of the 20 individual measurements taken from each Gateway $G_j$ (where j is the number of the Gateway at each location) at each reference location is used as the input to a fingerprint. The measure of the mean RSSI is robust in the presence of outlier values [91] and is computed for packets/RSSI (r) $i$ to $n$ in $G_j$ as in Equation 4.2;

$$RSSI\ mean\ for\ l\ = \frac{\sum_{i=1}^{n=20} r_i^{G_j}}{n} \qquad \textbf{Equation 4.2}$$

Where $l$ is the location measurment. Therefore, the mean RSSI values is used as location fingerprints.

## 4.6 Machine Learning Algorithms

### *4.6.1 K-Nearest Neighbour*

K-Nearest Neighbour (KNN) has been proven to be one of the simplest and most widely used algorithms in location fingerprinting [92] [93] [94]. KNN is applied as a regression problem that maps signal input features (RSSI) onto dual outputs representing location coordinates (longitude and latitude). The KNN algorithm relies on the assumption of locality in the feature space.

The KNN training phase is equivalent to creating a database (radio map) with reference LoRaWAN location patterns. The location pattern consists of a known location and a feature value. In this case, the feature value is the Received Signal Strength Indicator (RSSI) and each RSSI-location pair constitutes a training data point for the algorithm.

In the location estimation phase, the dataset composed RSSI values only. The distance between RSSI values and each stored neighbour is calculated in feature sub-space using the Euclidean distance metric [95] taking into account $k$ nearest neighbours with the shortest distances, as in Equation 4.3.

$$d = \sqrt{\sum \left(RSSI_{test} - RSSI_{training}\right)^2} \qquad \textbf{Equation 4.3}$$

The unknown LoRaWAN location coordinate is then calculated as the mean of the coordinates of the $k$ nearest neighbours, as in Equation 4.4.

$$(\widehat{lng}, \widehat{lat}) = \frac{1}{N}\sum_{i=1}^{N}(lng_i, lat_i) \qquad \textbf{Equation 4.4}$$

### *4.6.2 Weighted k-Nearest Neighbour*

The KNN can be modified such that the selected neighbours used in the final prediction are given weights leading to the Weighted k-Nearest Neighbour (WKNN) algorithm. The modification is proven to smooth out outliers. Neighbours closest to the query location are weighted more hence contributing more to the final prediction than further away neighbours, which potentially improves the model. The computed distances in feature space are often used to calculate the weight in each instance.

A variety of weight functions can be used [96, 97]. However, here, a simple weight function based on the inverse of the feature distance is adopted and will be used in subsequent analysis (Equation 4.5).

$$Weight = \frac{1}{d} \qquad d = \text{ is the Euclidean distance} \qquad \textbf{Equation 4.5}$$

It is important to note that in the machine learning technique, separate datasets are needed to train and validate the model. Here, the RSSI/location data collected during from 150 locations is randomly divided into training and test sets. A total of 120 x 4 (Gateways) randomly selected RSSIs with reference locations are used for training the model and 30 x 4 remaining RSSIs without reference locations to validate the developed models. The value of k affects the performance of KNN algorithm; the optimal value of k is determined using cross-validation.

## 4.7 Performance Analysis

### *4.7.1 Performance Metrics*

The performance of the developed node localisation models is evaluated based on the following metrics:

a) Accuracy

Accuracy is the deviation of the estimated location from the ground truth location of a node. The accuracy of the models is measured as the mean of the Haversian distance metric between the estimated location and the true location of a node given in Equation **4.6**

$$d = 2r \sin^{-1}\left(\sqrt{\sin^2\left(\frac{\varphi - \varphi_0}{2}\right) + \cos(\varphi_0)\cos(\varphi)\sin^2\left(\frac{\lambda - \lambda_0}{2}\right)}\right) \quad \textbf{Equation 4.6}$$

where, $\varphi_0 =$ latitude of real location

$\varphi =$ latitude of estimated location

$\lambda_0 =$ longitude of real location

$\lambda =$ longitude of estimated location

b) Cumulative Distribution Function (CDF)

The CDF determines the probability that the localisation error is less than or equal to a certain user-defined value. For example, the CDF of localisation error X is defined as in Equation **4.7**.

$$F_X(x) = P(X \leq x), for\ all\ x \in R$$ **Equation 4.7**

c) RMSE

RMSE is defined as the square root of the mean squared difference between the estimated location and the true location of a node; the localisation error in terms of RMSE is given by Equation **4.8**;

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(d_i)^2}$$ **Equation 4.8**

where $d_i$, is the Haversian distance, the distance difference between real and estimated locations (Equation **4.6**). The RMSE gives a measurement of how the data concentrated around the line of the best fit.

### *4.7.2 Effect of k*

The parameter k impacts the performance of the KNN. Larger values of k, render the model less sensitive to noise and much smoother. Here, cross-validation technique [98] is employed to determine the optimal value of k because it results in less bias. In general, cross-validation is a resampling procedure used to evaluate machine learning models on a limited data sample. The procedure has a single parameter called v that refers to the number of groups that a given data sample is to be split into. As such, the procedure is often called v-fold cross-validation. This involves randomly dividing the set of observation into v folds, of approximately equal size. In tuning for the optimal value of k parameter in KNN, the common value to use for v is 5. That is 5-fold cross-validation. Figure 4.8 shows a diagrammatic

representation of the 5-fold cross validation method. The first fold is treated as a validation set, and the remaining v-1 folds are used to train the model. The general procedure is as follows:

1. Shuffle the dataset randomly.
2. Split the dataset into v folds
3. For each unique fold and unique value of k:
    a. Take the fold as a hold out or test dataset
    b. Take the remaining folds as a training dataset
    c. Train the model and evaluate on the test set
    d. Retain the testing accuracy and discard the model
4. Take the mean of these test accuracy as the accuracy of the sample.
5. Finally, the k value with the best test accuracy is selected as optimal.

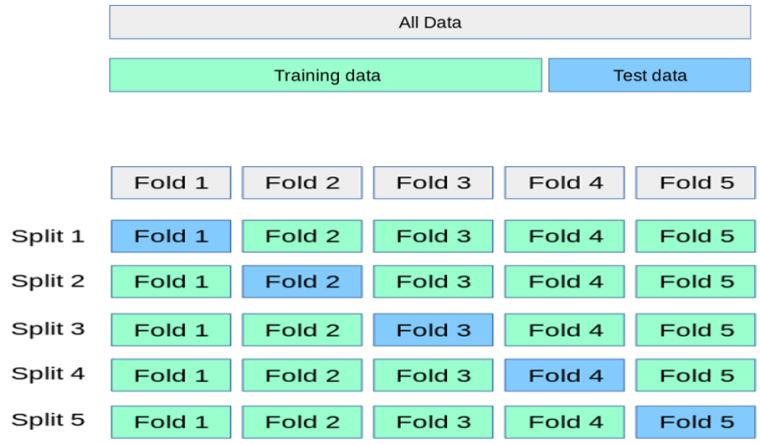



**Figure 4.8: *v*-fold cross-validation [99].**

A *5*-fold cross validation is used to determine the optimal k for each dataset collected for different spread factors (SF9, SF10, SF11, SF12). Figure 4.9 shown the effect of k on the

performance of the KNN for SF9, SF10, SF11 and SF12. The optimal value of k for the different datasets is presented in Table 4.5.

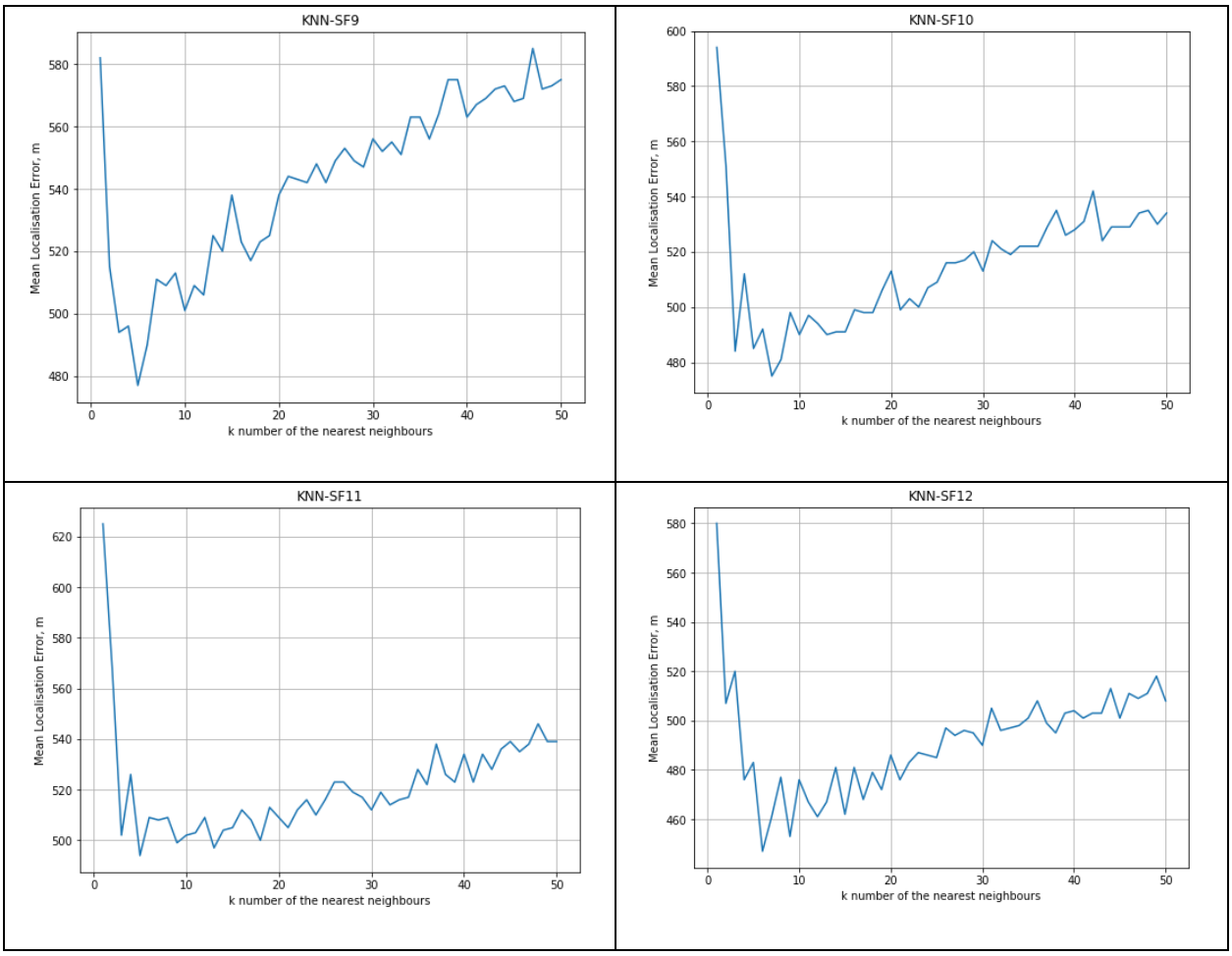



**Figure 4.9: K Graphs of kNN model for different spreading factors.**

**Table 4.5: Optimal k for different spreading factors.**

| Dataset | SF9 | SF10 | SF11 | SF12 |
|---|---|---|---|---|
| Optimal k | 5 | 7 | 5 | 6 |

### *4.7.3 Impact of Spreading Factor (SF)*

In order to evaluate the performance of KNN and WKNN used in locating node position, the data gathered detailed in Section 4.3.5 is used. The training data (120 x 4) and their corresponding ground truth locations are used to train the algorithms to establish the node localisation models for the sandstorm environment. The location estimation error is defined as the distance between the real location coordinates and the estimated location coordinates using the Haversian distance metric (Equation **4.6**).

The performance of the models is given in terms of their Cumulative Distribution Function (CDF) and statistical operators of location error [100]. The CDF describes the probability of locating the transmitter node within a localisation error range. Table 4.6 shows the statistical indices in the performance of each model on the different datasets based on the SF used. The comparison of the cumulative probability of localisation error between the models for different SF is presented in Figure 4.10.

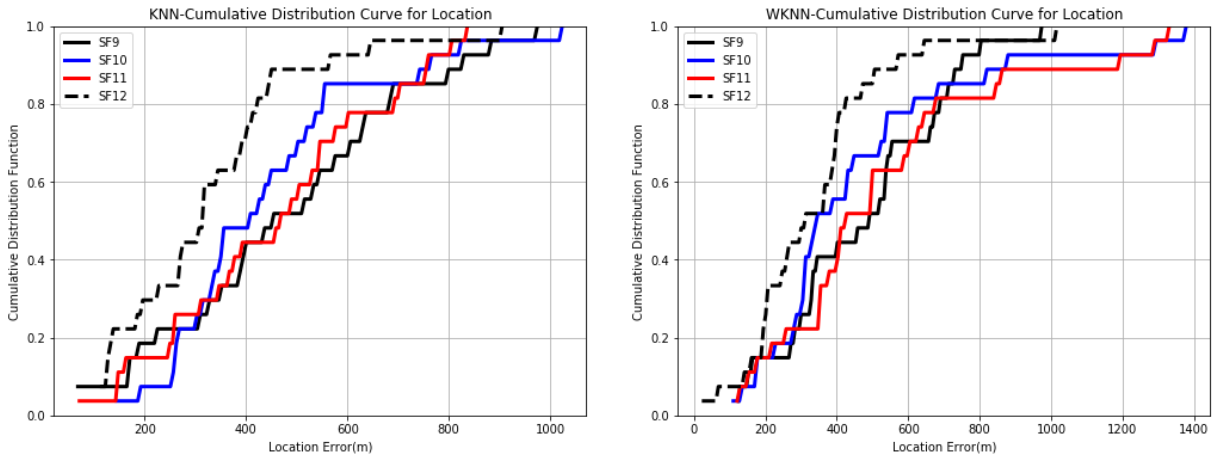



**Figure 4.10: Cumulative probability of localisation error for kNN and WkNN.**

**Table 4.6: Localisation performance of kNN and WkNN.**

| Model | SF9 (m) | SF10 (m) | SF11 (m) | SF12 (m) |
|---|---|---|---|---|
| **kNN (min)** | 65 | 142 | 68 | 109 |
| **kNN (max)** | 978 | 1017 | 839 | 908 |
| **kNN (mean)** | 476 | 440 | 459 | 326 |
| **kNN (median)** | 454 | 408 | 469 | 316 |
| **kNN (RMSE)** | 534 | 485 | 495 | 372 |
| **WkNN (min)** | 122 | 106 | 117 | 19 |
| **WkNN (max)** | 977 | 1382 | 1335 | 1019 |
| **WkNN (mean)** | 475 | 468 | 530 | 338 |
| **WkNN (median)** | 492 | 346 | 428 | 315 |
| **WkNN (RMSE)** | 525 | 561 | 622 | 391 |

It is evident from Table 4.6 that the localisation accuracy of the models improves as the spreading factor increases from 9 to 12. Specifically, the WKNN with SF12 can provide a median error in 315m, enhancing the precision of node localisation by 35.98 % over the performance of WKNN with SF9. Furthermore, WKNN with SF12 also provides a minimum localisation error of 19m. The trade-off between latency and accuracy in this application is a design factor e.g. at SF12 the node exhibit latency issues (delays in transfer of packets). At a low SF, shadowing and reflection will impact through reducing the reception at transmission node locations that would otherwise have been received at higher SFs.

### *4.7.4 Impact of Localisation Algorithm*

KNN and its weighted version, WKNN are used to model the RSSI-location relation and hence infer locations of new observed RSSIs. For each dataset harvested at different SFs during the foundation series of measurements, the optimal value of k was chosen empirically (Section 4.7.2). Table 4.6 summarises the performance comparison between KNN and WKNN in terms of RMSE, mean, median, maximum and minimum location error indicating that the localisation accuracy of the two models are comparable. The best mean localisation accuracy of KNN and WKNN is 326m and 338m, respectively. Although the minimum location error of KNN is inferior to that of WKNN, the maximum location error of KNN is less than that of WKNN in all cases.

## 4.8 Summary

An investigation into the use of RSSI based fingerprinting for LoRa node localisation in sand storm environments has been carried out. The mean RSSI at each location was used as the location feature.

Two machine learning algorithms, KNN and WKNN have been used to develop localisation models. An investigation into the impact of different spreading factors on node localisation performance has also been conducted.

The analysis conducted for both KNN and W-KNN is based on using the same distance measure (Euclidean) and same neigbour counts for all the datasets. Based on the obtained results, it is concluded that the performance of the models depends on the dataset. For dataset obtained using SF10, SF11and SF12, W-KNN shows improvement over KNN in terms of median localization error

with the least minimum error but fails to improve the overall performance. On the other hand, for SF9 dataset, KNN outperforms W-KNN. In both cases, the differences are not overwhelming. In this work, the expectation that W-KNN will improve results obtained by KNN was not achieved. This is due to the measured dataset, the constraint of distance measure and neighbor counts. At SF12, both models provide acceptable localisation accuracy in comparison to lower SFs (9, 10 and 11). The results contribute to informing on the feasibility of any localisation-based application for these environments.

The next Chapter will detail the use of new features derived from the RSSI data in conjunction with machine learning to improve node localisation performance.

# CHAPTER 5 KERNEL-BASED NODE LOCALISATION USING RSSI RATIOS

## 5.1 Introduction

In the previous Chapter, the use of RSSI as an input to develop simple machine learning models for localisation in sand storm environment was detailed. In this Chapter, more advanced machine learning algorithms and feature transformation is explored to enhance models for node location.

The RSSI values are transformed into ratios from pairs of Gateways. Two kernel based algorithms - Support Vector Regression (SVR) and Gaussian Process Regression (GPR) - are then used to model the relationship between the RSSI ratios and location; as in the previous Chapter, these models are then used to infer node locations.

Firstly, the location fingerprints are computed. The theoretical background of Gaussian Process Regression (GPR) and two versions of SVR, epsilon SVR and Nu SVR are discussed. Different kernel functions are used to developed several SVR and GPR models for node localisation. Finally, the Chapter concludes with a performance evaluation on the use of Kernel based methods in conjunction with RSSI ratios for LoRa node localisation in the selected environment.

## 5.2 RSSI Ratio Location Fingerprint

The highly unpredictable nature of absolute RSSI values used in location fingerprinting in environments characterised by reflections and obstructions introduces noise to the input

features and consequently adversely impacts the accuracy of localisation. Here, ratios of RSSI between pairs of receivers are used in order to limit the impact of the variations in the absolute RSSI values and provide more robust location fingerprints. The premise motivating the use of the RSSI ratio is that the location of the node can be uniquely determined if there are more than three spatially separated receivers which observed the RSSI readings from that transmission node.

Assuming $G = \{g_1, g_2, \ldots, g_n\}$ is a set of Gateways deployed in the area under consideration and $L = \{l_1, \ldots, l_m\}$ represents the reference node locations. The location feature space, $l_i$ can then be represented by Gateways and measured RSSI values $r \in R$ where $R = \{r_1, r_2, \ldots, r_n\}$. The RSSI ratio is defined at each location for a unique pair of Gateways.

The received signal strength ratio for the gateways $g_i$ and $g_j$ can be computed for measurement taken at location $l$= [($g_i$; $r_i$); ($g_j$; $r_j$)] as in Equation 5.1;

$$\boldsymbol{RSSI_{ratio}(g_i, g_j)} = \frac{r_i}{r_j} \qquad \text{With } i < j \text{ for uniqueness.} \qquad \textbf{Equation 5.1}$$

Where $r$ is absolute RSSI.

The mean of the RSSI ratios for each location were computed as in Equation 5.2;

$$Mean\ RSSI_{ratio} = \frac{\sum_{i,j=1}^{n=20} \frac{r_i}{r_j}}{n} \qquad \textbf{Equation 5.2}$$

where $g_{i,j}$ denotes the number of unique pair of Gateways that measures the signal strength of the node at location $l_i$. Mean RSSI ratios will be used in the subsequent analysis. The proposed node localisation technique is shown in Figure 5.1.

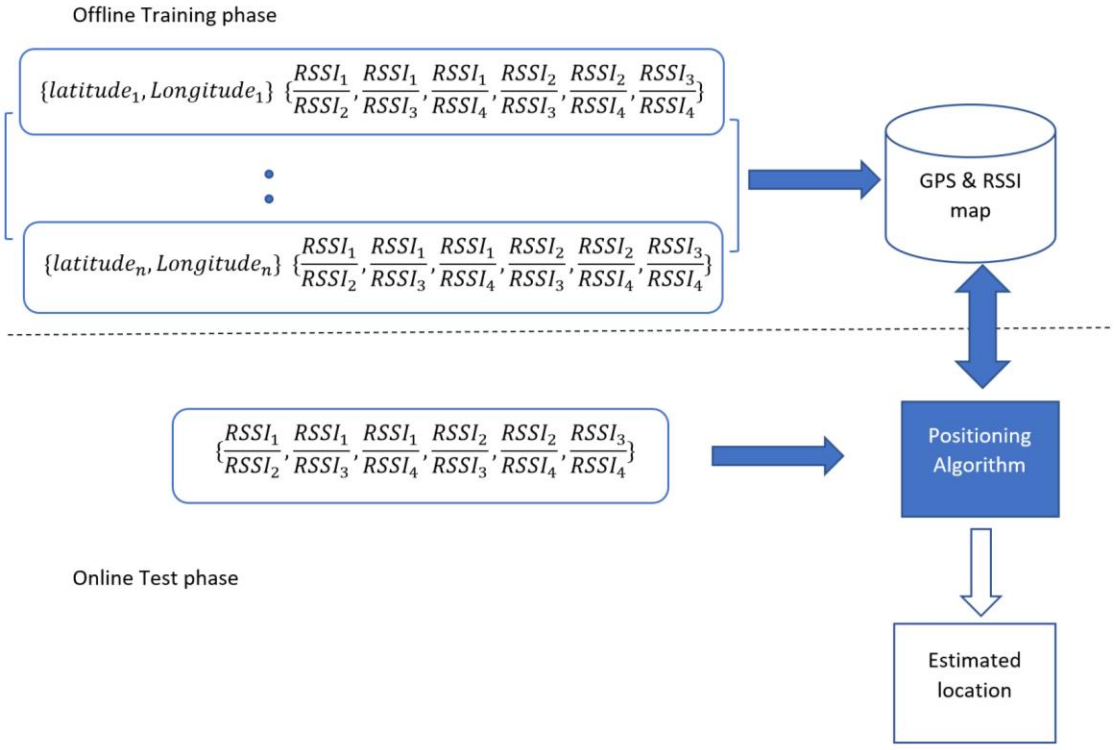



**Figure 5.1: Node localisation based on RSSI ratios.**

## 5.3 Kernel Methods

### *5.3.1 Support Vector Regression (SVR)*

SVR [101] [102] is a variant of the well-known Support Vector Machine (SVM) algorithm dedicated to regression problems. SVR is based on the same principles as SVM [103] for classification, using nonlinear mapping to transforming the data into a high dimensional feature space; linear regression is then executed in this space. The Kernel functions perform the nonlinear transform of the data into higher dimensional feature space that then enables the linear separation. The linear regression in a high dimensional space corresponds to a nonlinear regression in the low-dimensional input space [104].

Generally, regression methods derive a function (say) $f(\boldsymbol{x})$ with the least deviation between predicted and observed data for all training data. In SVR, the goal remains the same; to minimise the deviation or error, as shown in Figure 5.2. Given that the SVR output is a real number, it becomes very difficult to derive a prediction. Consequently, a tolerance margin epsilon is set in an approximation to the SVM.

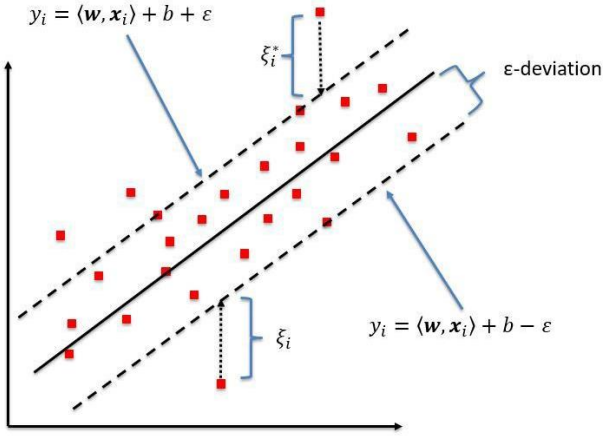



**Figure 5.2: Schematic of the one-dimensional Support Vector Regression (SVR) model [105].**

Two basic types of SVR are explored: epsilon-SVR and nu-SVR [106] [107], differing in the manner the parameters therein are controlled. In epsilon-SVR, no control on how many data vectors from the dataset become support vectors is invoked. Nonetheless, total control of how much error is allowed and anything beyond the specific epsilon is penalised in proportion to C, the regularisation parameter. On the other hand, nu-SVR determines the proportion of the number of support vectors with respect to the total number of samples in the dataset. In other words, nu represents an upper bound on the proportion of training samples that are errors and a lower bound on the proportion of samples that are support vectors. In nu-SVR, the parameter epsilon is introduced into the optimisation problem

formulation and is automatically estimated. The linear case of SVR is modeled as shown in Equation 5.3;

$$f(\boldsymbol{x}) = \langle \boldsymbol{w} \cdot \boldsymbol{x} \rangle + b \qquad \textbf{Equation 5.3}$$

The SVR problem can be written as a convex optimisation as stated in Equation 5.4 [108];

$$\text{Minimise } \frac{1}{2}\|\boldsymbol{w}\|^2 \qquad \textbf{Equation 5.4}$$

$$\text{Subject to } \begin{cases} y_i - \langle \boldsymbol{w} \cdot \boldsymbol{x_i} \rangle - b \leq \varepsilon \\ \langle \boldsymbol{w} \cdot \boldsymbol{x_i} \rangle + b - y_i \leq \varepsilon \end{cases}$$

ε – is the acceptable deviation of estimated locations from actual location.

An implicit assumption is that the function $f(\boldsymbol{x})$ can approximate all input pairs $(\boldsymbol{x_i}, y_i)$ with $\varepsilon$ precision, i.e. it is assumed optimisation is feasible. In order to accommodate errors, slack variables $\xi_i, \xi_i^*$ are introduced to cope with otherwise infeasible optimisation constraints giving Equation 5.5 [109], where the constant $C > 0$ determines the degree to which deviations larger than $\xi$ are tolerated with $l$ being the number of samples;

$$\text{Minimise } \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{l}(\xi_i + \xi_i^*) \qquad \textbf{Equation 5.5}$$

$$\text{Subject to } \begin{cases} y_i - \langle \boldsymbol{w} \cdot \boldsymbol{x_i} \rangle - b \leq \varepsilon + \xi_i \\ \langle \boldsymbol{w} \cdot \boldsymbol{x_i} \rangle + b - y_i \leq \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \geq 0 \end{cases}$$

$\xi_i, \xi_i^*$ - The slack variables make allowance for the localisation errors to exist up to the value of $\xi_i$ and $\xi_i^*$ without degrading performance. C - is the box constraint, a positive numeric value that controls the penalty imposed on data points that lie outside the ε margin and helps to prevent overfitting.

A standard dualisation method with Lagrange multipliers $\alpha_i, \alpha_i^*$ can be used [110] to solve the problem in Equation 5.5 and $\boldsymbol{\omega}$ can be expanded to Equation 5.6 [104];

$$\boldsymbol{w} = \sum_{i=1}^{l}(\alpha_i - \alpha_i^*)\boldsymbol{x}_i \text{ ; Where } \alpha_i \geq 0 \text{ and } \alpha_i^* \geq 0 \qquad \textbf{Equation 5.6}$$

Substituting Equation 5.6 into Equation 5.3 and Equation 5.5 produces Equation 5.7 [111];

$$f(\boldsymbol{x}) = \sum_{i=1}^{l}(\alpha_i - \alpha_i^*)\langle \boldsymbol{x_i} \cdot \boldsymbol{x} \rangle + b \qquad \textbf{Equation 5.7}$$

A variant of Equation 5.7 can be applied to develop nonlinear solutions by replacing the *dot* product of the input vectors with their nonlinear transformation, known as the Kernel function, represented by $k(\boldsymbol{x}_i, \boldsymbol{x})$ as in Equation 5.8 [111];

$$f(\boldsymbol{x}) = \sum_{i=1}^{l}(\alpha_i - \alpha_i^*)\, k(\boldsymbol{x}_i, \boldsymbol{x}) + b \qquad \textbf{Equation 5.8}$$

The use of Kernel functions makes SVR applicable to both linear and nonlinear approximations.

SVR gives a reasonable generalisation performance because it uses only the support vectors for prediction and is based on structural risk minimisation that seeks to minimise the generalisation rather than the training error [112].

### *5.3.2 Gaussian Process Regression (GPR)*

The Gaussian process is a probabilistic Kernel based machine learning technique that has been applied in many practical problems including estimation, classification, prediction and

prognosis due to its advantages of being flexible, probabilistic, and non-parametric [73] [106]. A Gaussian Process (GP) can model any system or process according to a normal or Gaussian distribution, where the mean and covariance function depends on the training data; the process is a collection of random variables with a joint Gaussian distribution [113]. Any function sample from the GP has a Gaussian distribution defined by its mean function $m(x)$ and covariance function $k(x, x')$.

The GP model assumes that the output is a realisation of a GP with joint probability density function given as Equation 5.9 [73];

$$f(x) \sim GP(m(x), k(x, x')) \qquad \textbf{Equation 5.9}$$

where, $m(x) = E(f(x))$

$$k(x, x') = E[\big(f(x) - m(x)\big)\big(f(x') - m(x')\big)]$$

$$k(x, x') = cov(f(x), f(x'))$$

Here, GP is applied to a regression problem.

Assume $X = [x_1, x_2, \ldots, x_N]$ represents N by 6-dimensional RSSI ratio input vectors, and the corresponding outputs are $y = [y_1, y_2, \ldots, y_N]$ representing the dual location coordinates. When a new input vector $x^*$ is given, the aim is to predict the corresponding output $y^*$ (location coordinates). The relationship between the input variable and the expected output can be modeled as Equation 5.10;

$$y_i = \varphi(x_i; W) + \varepsilon, \qquad \varepsilon \sim N(0, \sigma_n^2), \qquad i = 1, \ldots, N \qquad \textbf{Equation 5.10}$$

where $\varphi$ is a function parameterized by vector $W$; $\varepsilon$ is assumed to be the noise caused by disturbances or distributed Gaussian distribution N with zero mean and variance $\sigma_n^2$.

The prior probability on y is given by Equation 5.11 [106];

$$E[y] = E[\varphi(x; W) + \varepsilon] = 0 \qquad \textbf{Equation 5.11}$$

$$cov[y] = K(X, X) + \sigma_n^2 I$$

where E is the mean function, *cov* is the variance function.

The distribution with the new input can be expressed by Equation 5.12 [113];

$$\begin{bmatrix} y \\ y^* \end{bmatrix} \sim GP\left(0, \begin{bmatrix} K(X,X) + \sigma_n^2 I & K(X,x^*) \\ K(X,x^*)^T & K(x^*,x^*) \end{bmatrix}\right) \qquad \textbf{Equation 5.12}$$

where $K(X, x^*) = [k(x_1, x^*), \ldots, k(x_N, x^*)]$ can be written as $k^*$. The prediction can be presented by Equation 5.13 and Equation 5.14 [73];

$$E(y^*) = k^{*T}(K + \sigma_n^2 I)^{-1} y^T \qquad \textbf{Equation 5.13}$$

$$cov[y^*] = K(x^*, x^*) - k^{*T}(K + \sigma_n^2 I)^{-1} K^* \qquad \textbf{Equation 5.14}$$

## 5.4 Kernel Function

In machine learning, a Kernel is normally used to refer to a technique using a linear model to solve a nonlinear problem, implying the transformation of linearly inseparable to linearly separable data. The Kernel function is a function applied on each data point to map the original nonlinear observations into a higher dimensional space in which they become

separable. In simple terms, Kernel functions compute similarities between samples in the data.

A number of Kernel functions can be used in SVR and GPR algorithms. Here, commonly used Kernel functions are considered and new Kernels by combining kernel functions derived. The list of Kernel functions used are given in Table 5.1 [113] [114] [115].

**Table 5.1: Common Kernel functions [113] [114] [115].**

| | | |
|---|---|---|
| **1** | Linear | $k(\boldsymbol{x}, \boldsymbol{x}') = \boldsymbol{x}^{\boldsymbol{T}} \boldsymbol{x}'$ |
| **2** | Polynomial | $k(\boldsymbol{x}, \boldsymbol{x}') = \langle \boldsymbol{x} \cdot \boldsymbol{x}' \rangle^{d}$ |
| **3** | Radial Basis Function (RBF) | $k(\boldsymbol{x}, \boldsymbol{x}') = e^{\left(-\frac{\|x-x'\|^2}{2\sigma^2}\right)}$ |
| **4** | Sigmoid | $k(\boldsymbol{x}, \boldsymbol{x}') = \tanh(\gamma \boldsymbol{x}^{\boldsymbol{T}} \cdot \boldsymbol{x}')$ |
| **5** | Rational Quadratic (RQ) | $k(\boldsymbol{x}, \boldsymbol{x}') = (1 + \frac{d(\boldsymbol{x}, \boldsymbol{x}')^2}{2\alpha l^2})^{-\alpha}$ |
| **6** | Matern | $k(\boldsymbol{x}, \boldsymbol{x}') = \sigma^2 \frac{1}{\Gamma(\nu) 2^{\nu-1}} (\gamma\sqrt{2\nu d}(\frac{\boldsymbol{x}}{l}, \frac{\boldsymbol{x}'}{l}))^{\nu} \, k_{\nu}(\gamma\sqrt{2\nu d}(\frac{\boldsymbol{x}}{l}, \frac{\boldsymbol{x}'}{l}))$ |
| **7** | ExpSineSquared | $k(\boldsymbol{x}, \boldsymbol{x}') = \exp(-2(\sin(\frac{\pi}{p} * d(\boldsymbol{x}, \boldsymbol{x}'))/l^2)^2)$ |

## 5.5 Performance Analysis

### *5.5.1 Performance Evaluation Metrics*

Performance evaluation metrics provide the basis for a comparison of the developed node localisation models. SVR and GPR models are evaluated and the comparison is based on the

following metrics: accuracy, Cumulative Distribution Function (CDF) and Root Mean Square Error (RMSE).

a) Accuracy

Accuracy represents the deviation of the estimated location from the ground truth location of a node. The accuracy of the models is measured as the mean of the Haversian distance metric between the estimated location and the true location of a node given in Equation 5.15;

$$d = 2r \sin^{-1}\left(\sqrt{\sin^2\left(\frac{\varphi - \varphi_0}{2}\right) + \cos(\varphi_0)\cos(\varphi)\sin^2\left(\frac{\lambda - \lambda_0}{2}\right)}\right) \quad \textbf{Equation 5.15}$$

where, $\varphi_0 = latitude\ of\ real\ location$

$\varphi = latitude\ of\ estimated\ location$

$\lambda_0 = longitude\ of\ real\ location$

$\lambda = longitude\ of\ estimated\ location$

b) Cumulative Distribution Function (CDF)

The CDF is used to determine the probability that the localisation error is less than or equal to a certain user defined value. For example, the CDF of localisation error X is defined by Equation 5.16;

$$F_X(x) = P(X \leq x), for\ all\ x \in R \quad \textbf{Equation 5.16}$$

c) RMSE

RMSE is defined as the square root of the mean squared difference between the estimated location and the true location of a node. The localisation error in terms of RMSE is given by Equation 5.17, where $d$ is defined in Equation 5.15;

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(d_i)^2} \qquad \textbf{Equation 5.17}$$

### *5.5.2 Parameter Tuning*

The hyper-parameters associated with the machine learning algorithms affect overall performance when applied to a particular dataset. It is therefore imperative that these hyper-parameters be properly tuned in order to build an optimal model for the problem. The hyper-parameters are tuned for each dataset (RSSI ratios of SF9, SF10, SF11, and SF12 data), with the optimal model hyper-parameters for one particular dataset will not be the optimum across all datasets.

The random search method is used to select the optimal parameters of epsilon-SVR, nu-SVR and GPR algorithms. A grid of hyper-parameters values are established and a random combination of the values selected to train the model; here for SVR algorithm, hyper-parameter C, regularisation constant, epsilon and nu for nu-SVR; for GPR, the only hyper-parameter to tune is alpha. Some Kernels such as Matern has a parameter that is also optimised.

Figure 5.3 to Figure 5.6 show the elbow curves for tuning the epsilon parameter in SVR. Detail curves for other parameters can be found in the Appendix. The summary of the optimal parameters used in each algorithm for each dataset is given in Table 5.2.

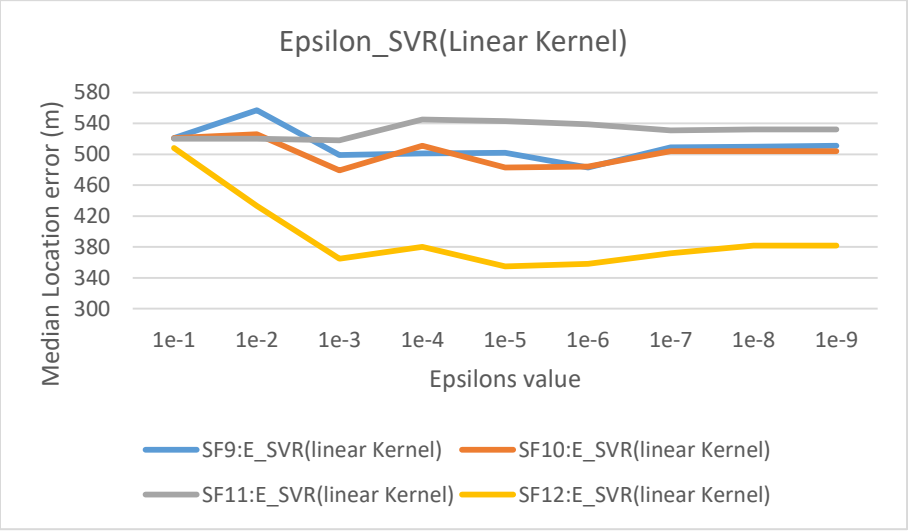



**Figure 5.3: Impact of epsilon on SVR performance using a linear Kernel.**

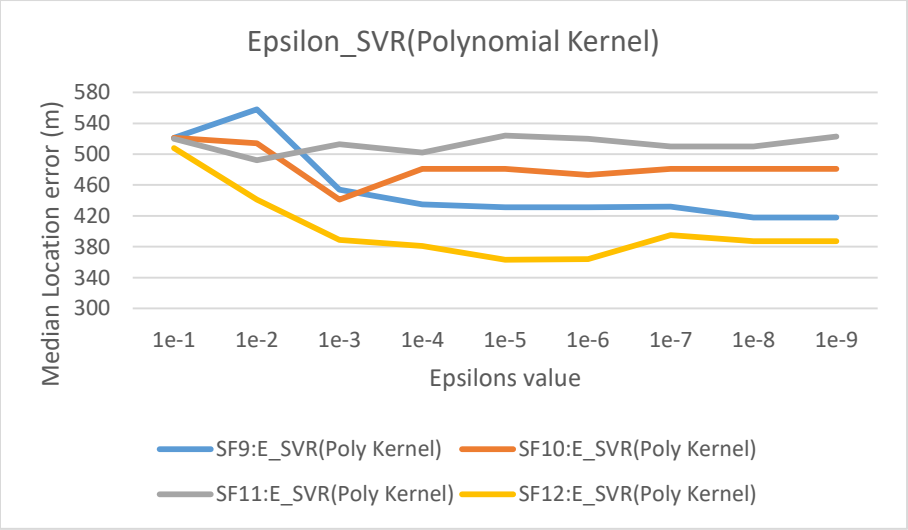



**Figure 5.4: Impact of epsilon on SVR performance using a polynomial Kernel.**

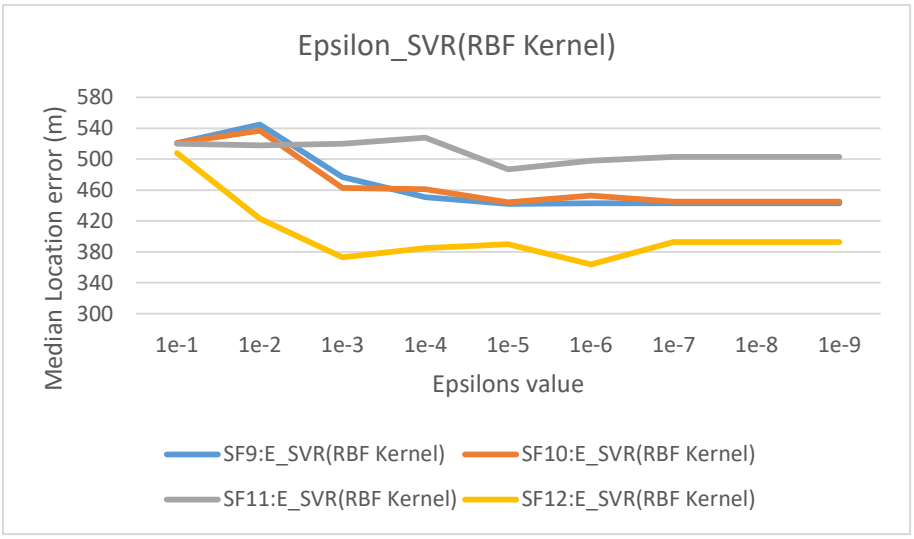



**Figure 5.5: Impact of epsilon on SVR performance using RBF Kernel.**

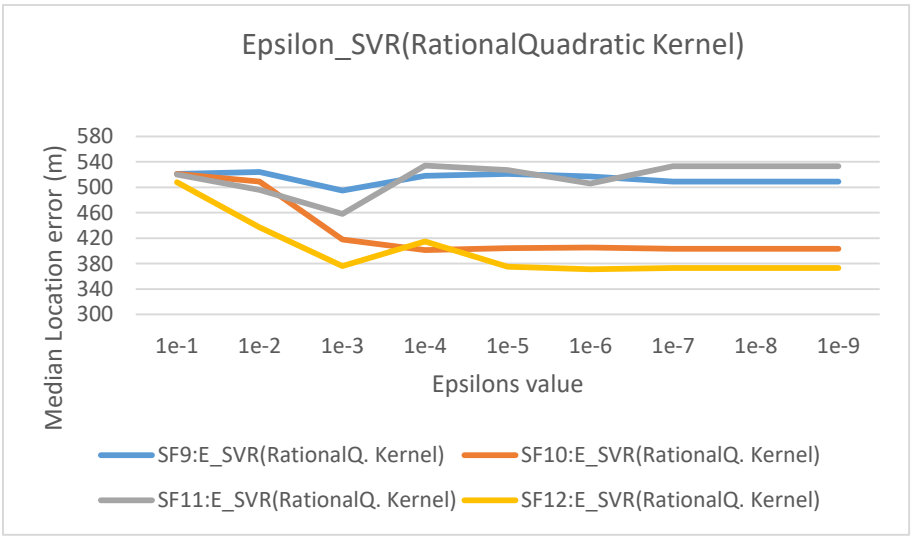



**Figure 5.6: Impact of epsilon on SVR performance using a rational quadratic kernel.**

**Table 5.2: Optimal parameters for each algorithm.**

| Kernels at GPR | SF 9_Ratio RSSI | | | SF 10_Ratio RSSI | | | SF 11_Ratio RSSI | | | SF 12_Ratio RSSI | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Alpha | Nu_Matern | median | Alpha | Nu_Matern | median | Alpha | Nu_Matern | median | Alpha | Nu_Matern | median |
| RBF+Matern | 1.00E-05 | 1.5 | 431 | 1.00E-06 | 8 | 378 | 1.00E-09 | 1 | 392 | 1.00E-07 | 9 | 317 |
| RQ+Matern | 1.00E-06 | 4 | 423 | 1.00E-08 | 11.5 | 388 | 1.00E-08 | 11 | 387 | 1.00E-07 | 10 | 361 |
| Exp.+Matern | 1.00E-05 | 2.5 | 430 | 1.00E-05 | 6 | 380 | 1.00E-09 | 1 | 392 | 1.00E-05 | 8 | 379 |

| Kernels at Nu_SVR | SF 9_Ratio RSSI | | | SF 10_Ratio RSSI | | | SF 11_Ratio RSSI | | | SF 12_Ratio RSSI | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Nu_SVR | Nu_Matern | median | Nu_SVR | Nu_Matern | median | Nu_SVR | Nu_Matern | median | Nu_SVR | Nu_Matern | median |
| RBF+Matern | 0.44 | 2 | 410 | 0.97 | 3 | 357 | 0.75 | 1 | 336 | 0.55 | 1 | 320 |
| RQ+Matern | 0.45 | 7 | 399 | 0.27 | 11 | 353 | 0.39 | 1 | 338 | 0.86 | 1.5 | 324 |
| Exp.+Matern | 0.09 | 1 | 425 | 0.22 | 9 | 366 | 0.47 | 1 | 338 | 0.78 | 1.5 | 309 |

| Kernels at Epsilon_SVR | SF 9_Ratio RSSI | | | SF 10_Ratio RSSI | | | SF 11_Ratio RSSI | | | SF 12_Ratio RSSI | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Epsilon | Nu_Matern | median | Epsilon | Nu_Matern | median | Epsilon | Nu_Matern | median | Epsilon | Nu_Matern | median |
| RBF+Matern | 1.00E-03 | 10 | 444 | 1.00E-03 | 4 | 359 | 1.00E-03 | 1 | 313 | 1.00E-06 | 3 | 323 |
| RQ+Matern | 1.00E-04 | 1.5 | 453 | 1.00E-10 | 8.5 | 381 | 1.00E-03 | 1 | 303 | 1.00E-06 | 14.5 | 346 |
| Exp.+Matern | 1.00E-03 | 1 | 477 | 1.00E-05 | 8 | 410 | 1.00E-03 | 1 | 326 | 1.00E-03 | 8.5 | 329 |

### *5.5.3 Impact of Kernel Function*

The effect of Kernel function used (see Table 5.2) in transforming data in original space into the higher dimension on the performance of the algorithm used for localisation is investigated. The RSSI ratio data derived from measured RSSI values and the corresponding location coordinates summarized in Table 5.3 are used as training inputs to the algorithms. It is important to note that only data of SF11 and SF12 are the focus since the preceding analyses have indicated that their performance is superior. While the data used for training remain constant, the Kernel function is varied in order to test the impact of the Kernel on the performance of each algorithm.

**Table 5.3: Summary of RSSI ratio data.**

| Dataset | No. of train | No. of test | Dimension of RSSI-Ratio | Dimension of Location Coordinate |
|---|---|---|---|---|
| **RSSI Ratios** | 120 | 30 | 6 | 2 |

Result with different Kernel functions for each algorithm - epsilon-SVR, nu-SVR and GPR - is shown in Table 5.4, Table 5.5, and Table 5.6 respectively. It is evident that the combined Kernel functions outperform the commonly used Kernels on the same dataset for all the three algorithms. More specifically, Rational Quadratic + Matern Kernel has the lowest median error of 303m in the epsilon-SVR algorithm viz. the epsilon-SVR model locates node with error less than 303m for 50% of the time. For the ExpSineSquared + Matern Kernel in nu-SVR, the median location error is 309m. In GPR, the RBF + Matern Kernel function gives a median error of 317 m with mean location error above 400 m.

**Table 5.4: Performance of different kernels on epsilon-SVR.**

| Epsilon_SVR Kernels | SF 11_Ratio RSSI | | | | SF 12_Ratio RSSI | | | |
|---|---|---|---|---|---|---|---|---|
| | Min (m) | Median (m) | Mean (m) | RMSE (m) | Min (m) | Median (m) | Mean (m) | RMSE (m) |
| **RBF** | 124 | 487 | 511 | 579 | 50 | 364 | 421 | 482 |
| **Linear** | 150 | 518 | 528 | 583 | 55 | 355 | 421 | 485 |
| **Polynomial** | 84 | 492 | 540 | 605 | 35 | 363 | 420 | 481 |
| **Sigmoid** | 74 | 491 | 534 | 606 | 77 | 462 | 524 | 589 |
| **RationalQuadratic** | 119 | 458 | 502 | 579 | 94 | 371 | 388 | 439 |
| **Matern** | 78 | 318 | 448 | 564 | 21 | 346 | 403 | 461 |
| **ExpSineSquared** | 115 | 390 | 481 | 575 | 49 | 342 | 395 | 451 |
| **Ex.+Matren** | 84 | 326 | 453 | 588 | 97 | 329 | 393 | 447 |
| **RBF+Matern** | 84 | 313 | 453 | 583 | 42 | 323 | 385 | 440 |
| **RQ+Matern** | 85 | 303 | 451 | 573 | 70 | 346 | 378 | 431 |

**Table 5.5: Performance of different kernels on nu-SVR.**

| nu_SVR Kernels | SF 11_Ratio RSSI | | | | SF 12_Ratio RSSI | | | |
|---|---|---|---|---|---|---|---|---|
| | Min (m) | Median (m) | Mean (m) | RMSE (m) | Min (m) | Median (m) | Mean (m) | RMSE (m) |
| **RBF** | 96 | 464 | 519 | 582 | 44 | 358 | 415 | 476 |
| **Linear** | 132 | 492 | 526 | 588 | 52 | 351 | 422 | 497 |
| **polynomial** | 126 | 471 | 500 | 562 | 31 | 357 | 413 | 474 |
| **Sigmoid** | 72 | 459 | 541 | 621 | 101 | 465 | 528 | 593 |
| **RationalQuadratic** | 85 | 417 | 506 | 581 | 89 | 363 | 400 | 450 |
| **Matern** | 143 | 337 | 454 | 559 | 80 | 327 | 401 | 450 |
| **ExpSineSquared** | 139 | 392 | 484 | 555 | 146 | 336 | 399 | 449 |
| **RBF+Matern** | 129 | 336 | 454 | 567 | 83 | 320 | 396 | 443 |
| **RQ+Matern** | 109 | 338 | 463 | 555 | 121 | 324 | 391 | 444 |
| **Exp.+M** | 111 | 338 | 474 | 612 | 116 | 309 | 404 | 459 |

**Table 5.6: Performance of different kernels on GPR.**

| GPR Kernels | SF 11: Ratio_RSSI | | | | SF 12: Ratio_RSSI | | | |
|---|---|---|---|---|---|---|---|---|
| | Min (m) | Median (m) | Mean (m) | RMSE (m) | Min (m) | Median (m) | Mean (m) | RMSE (m) |
| **RBF** | 21 | 395 | 483 | 595 | 100 | 364 | 432 | 513 |
| **RationalQuadratic** | 51 | 418 | 427 | 485 | 91 | 394 | 435 | 490 |
| **Matern** | 30 | 387 | 519 | 700 | 114 | 379 | 444 | 534 |
| **ExpSineSquared** | 68 | 400 | 445 | 498 | 87 | 397 | 418 | 475 |
| **Exp.+Matern** | 32 | 392 | 488 | 595 | 62 | 379 | 424 | 481 |
| **RBF+Matern** | 32 | 392 | 488 | 595 | 49 | 317 | 425 | 491 |
| **RQ+Matern** | 35 | 387 | 490 | 678 | 22 | 361 | 385 | 433 |

### *5.5.4 Evaluating Model Accuracy*

In order to evaluate the performance of the three algorithms (epsilon-SVR, nu-SVR and GPR), the RSSI ratio data detailed in Section 5.2 was used in tandem with the combined (derived) Kernels owing to their superior performance. RSSI ratio features (120 X 6) and their corresponding location coordinates were used as inputs to train the algorithms; RSSI ratios from the remaining 30 locations were used as test data. The statistical measures of location error for each node localisation model using the three best Kernels are presented in Table 5.7, Table 5.8, and Table 5.9.

**Table 5.7: Statistical measures of location error for each model using Rational Quadratic + Matern Kernel Function.**

| Models | SF9 RSSI Ratio (m) | SF10 RSSI Ratio (m) | SF11 RSSI Ratio (m) | SF12 RSSI Ratio (m) |
|---|---|---|---|---|
| **Epsilon-SVR(min)** | 59 | 66 | 85 | 70 |
| **Epsilon-SVR (median)** | 453 | 381 | 303 | 346 |
| **Epsilon-SVR (mean)** | 571 | 451 | 451 | 378 |
| **Epsilon-SVR (RMSE)** | 694 | 509 | 573 | 431 |
| **Nu-SVR (min)** | 119 | 57 | 109 | 121 |
| **Nu-SVR (median)** | 399 | 353 | 338 | 324 |
| **Nu-SVR (mean)** | 508 | 440 | 463 | 391 |
| **Nu-SVR (RMSE)** | 575 | 503 | 555 | 444 |
| **GPR(min)** | 49 | 52 | 35 | 22 |
| **GPR(median)** | 423 | 388 | 387 | 361 |
| **GPR(mean)** | 489 | 447 | 490 | 385 |
| **GPR(RMSE)** | 560 | 506 | 678 | 433 |

**Table 5.8: Statistical measures of location error for each model using ExpSineSquared + Matern Kernel Function.**

| Models | SF9 RSSI Ratio (m) | SF10 RSSI Ratio (m) | SF11 RSSI Ratio (m) | SF12 RSSI Ratio (m) |
|---|---|---|---|---|
| **Epsilon-SVR(min)** | 90 | 108 | 84 | 97 |
| **Epsilon-SVR (median)** | 477 | 410 | 326 | 329 |
| **Epsilon-SVR (mean)** | 637 | 498 | 453 | 393 |
| **Epsilon-SVR (RMSE)** | 767 | 559 | 588 | 447 |
| **Nu-SVR (min)** | 140 | 56 | 111 | 116 |
| **Nu-SVR (median)** | 425 | 366 | 338 | 309 |
| **Nu-SVR (mean)** | 541 | 432 | 474 | 404 |
| **Nu-SVR (RMSE)** | 609 | 498 | 612 | 459 |
| **GPR(min)** | 36 | 59 | 32 | 62 |
| **GPR(median)** | 430 | 380 | 392 | 379 |
| **GPR(mean)** | 503 | 450 | 488 | 424 |
| **GPR(RMSE)** | 576 | 508 | 595 | 481 |

**Table 5.9: Statistical measures of location error for each model using RBF + Matern Kernel Function.**

| Models | SF9 RSSI Ratio (m) | SF10 RSSI Ratio (m) | SF11 RSSI Ratio (m) | SF12 RSSI Ratio (m) |
|---|---|---|---|---|
| **Epsilon-SVR(min)** | 75 | 45 | 84 | 42 |
| **Epsilon-SVR (median)** | 444 | 359 | 313 | 323 |
| **Epsilon-SVR (mean)** | 532 | 449 | 453 | 385 |
| **Epsilon-SVR (RMSE)** | 629 | 507 | 583 | 440 |
| **Nu-SVR (min)** | 55 | 64 | 129 | 83 |
| **Nu-SVR (median)** | 410 | 357 | 336 | 320 |
| **Nu-SVR (mean)** | 518 | 452 | 454 | 396 |
| **Nu-SVR (RMSE)** | 600 | 514 | 567 | 443 |
| **GPR(min)** | 40 | 45 | 32 | 49 |
| **GPR(median)** | 431 | 378 | 392 | 317 |
| **GPR(mean)** | 502 | 454 | 488 | 425 |
| **GPR(RMSE)** | 572 | 526 | 595 | 491 |

Results indicate a consistency in the performance of the algorithms used for SF11 and SF12 irrespective of the combined Kernels. More specifically, each of the models provide a localisation accuracy with median error less than 400m, evident in the box plot for each model as a function of different Kernels in Figure A4.1 to Figure A4.9 in Appendix 4. Epsilon-SVR has the lowest median error of 303m compare to 309m and 317m for nu-SVR and GPR respectively. Furthermore, the overall performance of the three models is captured by CDFs

of the localisation error as shown in Figure A4.1 to Figure A4.9 in Appendix 4. SVR (303m) outperformed the GPR (317m) models in terms of overall accuracy. 82% of the time both nu-SVR and epsilon-SVR attempt to locate node with localisation error of 600m or less compared to GPR at 78%.

## 5.6 Conclusions

An investigation into the use of Kernelised learning methods with RSSI ratios for node localisation has been detailed. Specifically, epsilon- and nu-Support Vector Regression and Gaussian Process Regression have been used to model the complex relationship between RSSI ratios and node location. The RSSI ratio pairs are inputs to the models during training. The performance of the models has been evaluated and results indicate that the combination of different Kernel functions can enhance localisation accuracy.

The kernel-based models provide consistent performance when used with RSSI data at SF11 and SF12 irrespective of the combined Kernels. More specifically, each of the models provide a localisation accuracy with median error less than 400m. Epsilon-SVR has the lowest median error of 303m compare to 309m and 317m for nu-SVR and GPR respectively.

In the next chapter, a location fingerprint combination technique based on SFs will be explored and ensemble machine learning algorithm based on decision regression trees - random forest and gradient boosting - will be used to develop a node localisation model based on the combined features. Based on preceding results, the motivation is that a combination of features from different SFs provide a more robust set of features for localisation and the ensemble methods will improve accuracy.

# CHAPTER 6 NODE LOCALISATION VIA FEATURE COMBINATION AND ENSEMBLE METHODS

## 6.1 Introduction

The Chapter investigates the feasibility of using machine a learning ensemble technique and feature engineering to improve the performance of node localisation. To this end, RSSI parameters (RSSI values and ratios) at different Spreading Factors are combined to form a location fingerprint. Preceding analyses informs that a combination of location features based on SF enhance localisation accuracy. Therefore, a number of SF combinations are used to form different features that are used as input to develop models. Furthermore, machine learning ensemble methods - Random Forest (RF) and Gradient Boosting Regressor (GBR) - derived from Regression Decision Trees (RDTs) are investigated and used to model the complex relationship between the combined features and the location of nodes. The models are then used to infer node location. The performance of the node localisation models are evaluated with particular focus on accuracy and on comparison with RDT.

## 6.2 Location Fingerprint based on Combined Spreading Factors

Preceding analyses has confirmed that the Spreading Factor impacts the RSSI data collected, which has motivated the evaluation on the use of a combination of two different spread factors viz. RSSI and RSSI ratio as a combined feature for node localisation. The combined features may capture the varying degrees of attenuation, interference and the impact of the challenging environmental characteristics; therefore, With SF11 and SF12, established to produce the best predictions, there the hypothesis is that features that result from the

combination of spreading factors and ratios may be potentially more robust as location fingerprints.

Given the data from spreading factors SF9, SF10, SF11 and SF12 reported in Chapter 3, an observation for a given location can be represented by SF, nodes and measured RSSI. Therefore, a combination of RSSI features at each location for unique SF pair $SF_i$ and $SF_j$ given as $\{(RSSI_1^i, RSSI_2^i, RSSI_3^i, RSSI_4^i)( RSSI_1^j, RSSI_2^j, RSSI_3^j, RSSI_4^j)\}$ can be formulated (Figure 6.1). Similarly, a combination of RSSI ratios feature $\{(RSSI_1^i, RSSI_2^i, RSSI_3^i, RSSI_4^i, RSSI_5^i, RSSI_6^i)( RSSI_1^j, RSSI_2^j, RSSI_3^j, RSSI_4^j, RSSI_5^j, RSSI_6^j)\}$ - as explained in Chapter 5 - where i,j = 9, 10, 11, 12 represent spreading factors can be established. It is envisaged that this combination will increase the uniqueness of the RSSI values mapped to each known location, yielding more distinct fingerprints.

For the purpose of the evaluation, the combined features (both RSSI and RSSI ratios) based on SFs will be referred to as SF9 & SF10, SF9 & SF11, SF9 &SF12, SF10 & SF11, SF10 & SF12 and SF11 & SF12, indicating the SF-RSSI values used in each unique combination. The formulation can be extended to a combination of three or four SFs viz. SF9&SF10&SF11, SF9&SF10&SF12, SF9&SF11&SF12 or SF9&SF10&SF11&SF12. This nomenclature is used throughout the remainder of the Chapter. Table 6.1 is a summary of data sets generated by feature combination for different SFs.

**Table 6.1: Datasets from combined RSSI /RSSI ratio features.**

| Dataset | No. train | No. test | RSSI Dimension | RSSI_Ratios Dimension |
|---|---|---|---|---|
| **2 SFs Combined** | 120 | 30 | 8 | 12 |
| **3 SFs Combined** | 120 | 30 | 12 | 18 |
| **4 SFs Combined** | 120 | 30 | 16 | 24 |

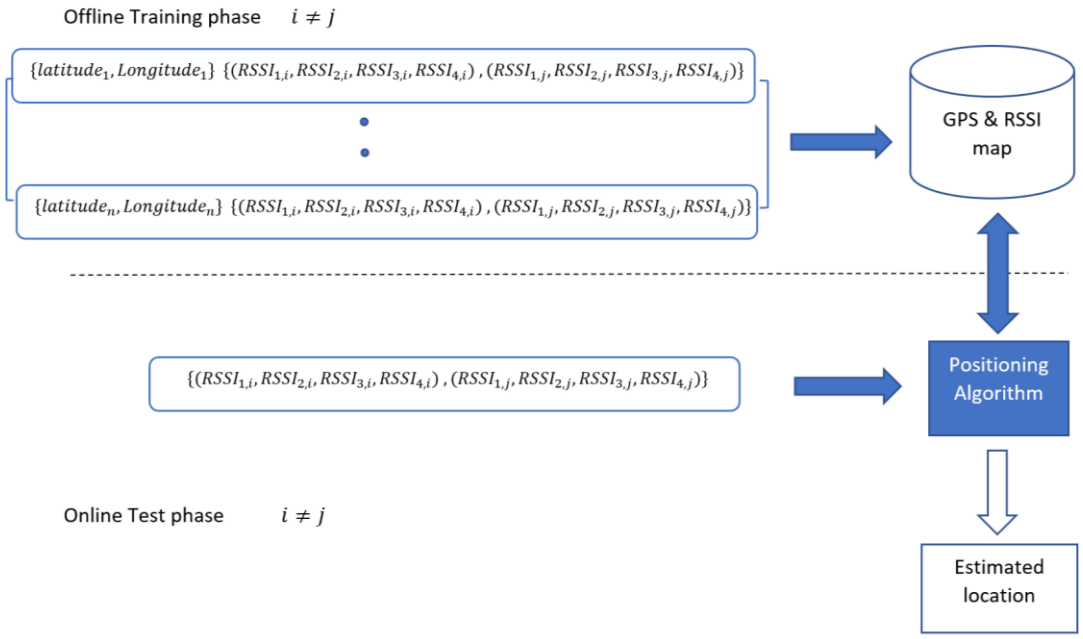



**Figure 6.1: Node localisation methodology based on combined features.**

## 6.3 Feasibility of using combined features

To verify the feasibility of using combined features based on SF, the KNN algorithm in Chapter 3 is used as a benchmark to model the combine RSSI-location relation. The combined features and their reference locations are used as KNN inputs to train and establish a model.

The KNN search space remains the same irrespective of features. K-nearest neighbours from the database are selected and used to estimate the location of a new observation. Since the combined features differ from the features used in the KNN developed in Chapter 3, the optimal value of parameter k derived is no longer valid. Therefore, an optimal value of k for combined features needs to be computed. V-fold cross validation is applied again on the training portion of the data to estimate the optimal value of the k-parameter. The results of v-fold cross validation for the optimal value of k in each pair of combine features (SF9 & SF10, SF9 & SF11, SF9 &SF12, SF10 & SF11, SF10 & SF12 and SF11 & SF12 is summarised in Table 6.2.

**Table 6.2: Optimal value of k for each data set.**

| | SF9&SF10 | SF9&SF11 | SF9&SF12 | SF10&SF11 | SF10&SF12 | SF11&SF12 |
|---|---|---|---|---|---|---|
| **Optimal_K** | 8 | 15 | 3 | 19 | 6 | 10 |

The optimal values of k are then used in the KNN model to estimate the locations of the nodes in the test data. The results of the evaluations are summarised in Table 6.3. A significant improvement in performance compared with the results with single SF (Chapter 3) is evident. Specifically, the best median localisation error in the combined data set is 297m for SF9&SF12, an improvement of 6% from the best median localisation error (316m) when single SF data is used. In conclusion, the result demonstrate that the technique of combining features from different SFs produces a manifest improvement in localisation accuracy.

**Table 6.3: Performance of kNN with two SFs feature combinations.**

| | SF9&SF10 (m) | SF9&SF11 (m) | SF9&SF12 (m) | SF10&SF11 (m) | SF10&SF12 (m) | SF11&SF12 (m) |
|---|---|---|---|---|---|---|
| **kNN(min)** | 43 | 140 | 99 | 91 | 85 | 35 |
| **kNN(max)** | 1008 | 1048 | 857 | 986 | 1066 | 968 |
| **kNN(mean)** | 469 | 496 | 322 | 418 | 377 | 365 |
| **kNN(median)** | 472 | 487 | 297 | 389 | 358 | 372 |
| **kNN(RMSE)** | 539 | 545 | 378 | 472 | 425 | 419 |

## 6.4 Regression Decision Trees (RDTs)

Decision Trees (DTs) are one of the simplest and widely used machine learning algorithms for classification and prediction [116] [117]. Regression Decision Trees (RDTs) [118, 119, 120] are a variant of the DT where the target variables take continuous values. Here, the DRT [121] [122] is regarded as a function approximation problem consisting of mapping of the node RSSI input onto output variables representing the latitude and longitude of the node position achieved through the training of the DRT to learn the complex relationship between the RSSI features and their respective referenced locations. The training/learning process involves splitting the whole data into smaller clusters handled by simple linear predictors. A regression model is established and in turn used to infer location of nodes given new feature observations.

The splitting process is achieved top-down from the root node to the leaves using recursive binary division. Recursive binary splitting is an iterative process that splits the training data (RSSI values and location co-ordinates) at each node into smaller groups with more

homogeneous/similar data points. In other words, the predictor space of possible feature-locations is divided into distinct and non-overlapping regions ($R$) as determined by Equation 6.1 and Equation 6.2:

$$\boldsymbol{R_1(j,s) = \{x | x_j \leq s\}} \quad \textbf{Equation 6.1}$$

$$\boldsymbol{R_2(j,s) = \{x | x_j > s\}} \quad \textbf{Equation 6.2}$$

Where, $j =$ splitting variable

$s =$ split point

For every observation that falls into a region, a prediction is made i.e. the mean of location coordinates in the training set in that particular region as determined Equation 6.3 and Equation 6.4:

$$(\widehat{lng}, \widehat{lat})_1(j,s) = ave\{(lng, lat)_i | x_i \in R_1(j,s)\} \quad \textbf{Equation 6.3}$$

$$(\widehat{lng}, \widehat{lat})_2(j,s) = ave\{(lng, lat)_i | x_i \in R_2(j,s)\} \quad \textbf{Equation 6.4}$$

where $j$ and $s$ are found by minimising loss ($L$) through a sum of Mean Square Error in partitioning the data as defined in Equation 6.5;

$$\boldsymbol{L(j,s)} = \sum_{i:x_i \in R_1(j,s)} \frac{1}{N_{R_1}} \left( (lng, lat)_i - \left(\widehat{lng}, \widehat{lat}\right)_{R_1} \right)^2 + \sum_{i:x_i \in R_2(j,s)} \frac{1}{N_{R_2}} \left( (lng, lat)_i - \left(\widehat{lng}, \widehat{lat}\right)_{R_2} \right)^2 \quad \textbf{Equation 6.5}$$

Location coordinates are subsequently inferred from new RSSI values inputted into the DT model that lead to a particular output leaf node. The tree expands until all leaves contain less than the minimum number of samples required to split an internal node.

A known problem with DRT is that there is the tendency to over-fit data. Although the model performs well on the training data, it is likely to have a greater error rate with unseen data [123]. In order to overcome overfitting, ensemble learning methods are employed. In the next section, the ensemble methods considered are discussed.

## 6.5 Ensemble Methods

The two ensemble methods employed in this thesis are GBR and RF. In machine learning, these models combine the results from multiple models to improve the performance in terms of prediction accuracy as compared to using a single model. The main causes of error in learning models are due to noise, bias and variance. Ensemble methods help to minimize these factors.

### *6.5.1 Random Forest (RF) for node localisation*

Random Forests (RFs) are combination of different individual tree models such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest [124, 125, 126]. The major aspect of RF is that its component trees are randomised in order to de-correlate individual tree predictions, in turn leading to improved generalisation and robustness. In other words, each tree of the forest is trained on a random subset of the training data. Both the samples and the features are

randomly selected with replacement. RFs have been used for classification and regression in many areas of studies [127, 128, 129]. Figure 6.2 shows the schema of the RF technique.

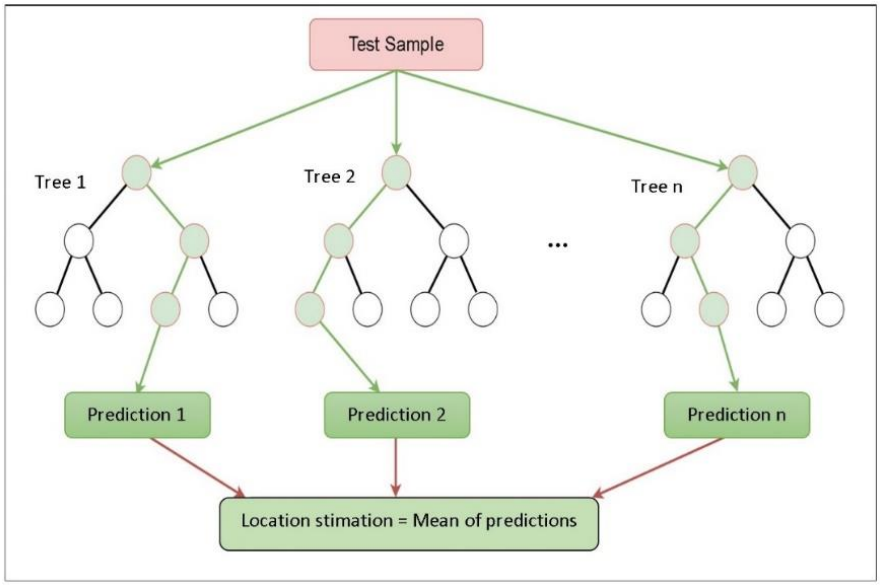



**Figure 6.2: Random forest procedure.**

In the node localisation application, the RF maps RSSI combined features to their location coordinates. The RF algorithm can be divided into an off-line training phase and an on-line test phase. In the training phase, individual tree models are built with bootstrap samples of the data. After obtaining multiple tree models, each testing sample is simultaneously pushed through all trees in multiple models until it reaches the corresponding leaves. RF test is executed in parallel, thus achieving high computational efficiency.

### *6.5.2 Gradient Boosting Regression (GBR)*

Gradient Boosting [130] [131] [132] is a variant of the Boosting machine learning ensemble technique for regression based on weighting observations, placing more weight on difficult

to predict cases and less on those well predicted. The result is a prediction model in the form of an ensemble of regression decision trees. In a statistical framework setting, boosting can be interpreted as an optimisation problem where the objective is to minimise the loss of the model by adding weak learners (decision trees) using a gradient descent like procedure [133]. The model is built in a stage-wise manner by adding one new regression tree at a time to compensate the shortcomings (minimise the loss) of existing weak regression trees in the model, whilst retaining existing models [134]. Therefore, GBR involves three elements; a loss function to be optimised, a weak learner (regression decision tree) to make predictions and an additive model to add weak regression tree learners to minimise the loss function. The loss function used is 'square loss', which implies that the task is to find an approximate model that minimises 'square loss' of the training data. In other words, using a training set $\{(x_i, y_i)\}_{i=1}^{n}$ of known values of RSSI, $x$ and corresponding values of location coordinates $y$, GBR tries to find an approximation $\hat{F}(x)$ that minimises the mean value of the loss function $L(y, F(x))$ on the training set (Equation 6.6);

$$L(y, F(x)) = \frac{1}{2}\big(y - F(x)\big)^2 \qquad \textbf{Equation 6.6}$$

Regression decision tree used as the weak learner in GB outputs real values for splits and those outputs can be added together, allowing subsequent models outputs to be added to compensate for the loss in the predictions. A gradient descent procedure is used to minimise the loss when adding trees [135]. Instead of minimising a set of parameters, the tree added to the model to reduce the loss is parameterised, and the parameters of the tree modified and moved in the right direction, reducing the error. The output of the new tree is then added

to the output of the existing sequence of trees in order to improve the final output of the model. Either a fixed number of trees are added or training stops once loss reaches an acceptable level or does not improve on an external validation dataset. Each $F_{m+1}(x)$ attempts to correct the errors of predecessor $F_m(x)$.

## 6.6 Performance Analysis

The data is randomly divided into two sets; the training set containing 120 data samples (RSSI) with corresponding location coordinates with the remaining 30 samples are taken as test or validation set.

### *6.6.1 Performance metrics*

The performance of the developed models is evaluated and compared based on the following metrics;

1. Cumulative Distribution Function (CDF)

The CDF metric is used to determine the probability that the localisation error is less than or equal to a certain user-defined value. For example, the CDF of localization error X is defined as Equation 6.7;

$$F_X(x) = P(X \leq x), for\ all\ x \in R \qquad \textbf{Equation 6.7}$$

2. Accuracy

Accuracy represents the deviation of the estimated location from the ground truth location of a node. The accuracy of the models is measured as the mean of the Haversian distance metric between the estimated location and the true location of a sensor given in Equation 6.8;

$$d = 2r \sin^{-1}\left(\sqrt{\sin^2\left(\frac{\varphi - \varphi_0}{2}\right) + \cos(\varphi_0)\cos(\varphi)\sin^2\left(\frac{\lambda - \lambda_0}{2}\right)}\right) \quad \textbf{Equation 6.8}$$

Where, $\varphi_0 = latitude\ of\ real\ location$

$\varphi = latitude\ of\ estimated\ location$

$\lambda_0 = longitude\ of\ real\ location$

$\lambda = longitude\ of\ estimated\ location$

3. RMSE

RMSE is defined as the square root of the mean squared difference between the estimated location and the true location of a node. The localisation error in terms of RMSE is given by Equation 6.9, where $d$ is defined in Equation 6.9;

$$\boldsymbol{RMSE} = \sqrt{\frac{\mathbf{1}}{\boldsymbol{n}}\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}}(\boldsymbol{d_i})^{\mathbf{2}}} \quad \textbf{Equation 6.9}$$

### *6.6.2 Impact of the Number of Trees and Maximum Depth*

Since the size of ensemble methods (RF and GB) changes as the number of trees utilised and maximum depth varies, it is important to examine the impact of both variables. In the

evaluation, other parameters such as number of features was not set to any value in order to generate each tree in a different state.

A number of individual RFs and GBs were constructed with different number of trees and different depths to evaluate the effect of the size of the ensemble methods. The effect of number of trees on the localisation accuracy is then assessed for each combination of features, representing different data sets. The combinations of two SFs, 3SFs and 4SFs, create eleven sets of data, with the evaluation assessing the impact of number of trees for each of the data sets. Figure 6.3, Figure 6.4, and Figure 6.5 show the so-called elbow curves indicating the performance of RF on 2SFs, 3SFs and 4SFs combined data respectively for different number of trees. Similar curves for RSSI ratios and the GB algorithm are shown in Figure 6.6 to Figure 6.14. A unique optimal number of trees with the lowest median location error each data combination is determined.

The summary of the optimal number of trees for each data combination (RSSI values and RSSI Ratios) is presented in Table 6.4. These optimal values of the number of trees are used in the remainder of the Chapter.

**Table 6.4: Optimal number of tree in RF and GB for different data combinations.**

| Combined data by SFs | Optimal number of trees | | | |
|---|---|---|---|---|
| | RF_RSSI | RF_Ratios | GB_RSSI | GB_Ratios |
| **SF9,SF10** | 66 | 21 | 24 | 156 |
| **SF9,SF11** | 3 | 18 | 16 | 24 |
| **SF9,SF12** | 13 | 3 | 23 | 63 |
| **SF10,SF11** | 12 | 9 | 54 | 58 |
| **SF10,SF12** | 8 | 29 | 13 | 17 |
| **SF11,SF12** | 20 | 84 | 75 | 139 |
| **SF9,SF10,SF11** | 21 | 8 | 108 | 34 |
| **SF9,SF10,SF12** | 18 | 3 | 76 | 28 |
| **SF9,SF11,SF12** | 4 | 49 | 109 | 167 |
| **SF10,SF11,SF12** | 11 | 5 | 191 | 40 |
| **SF9,10,11,12** | 26 | 6 | 112 | 184 |

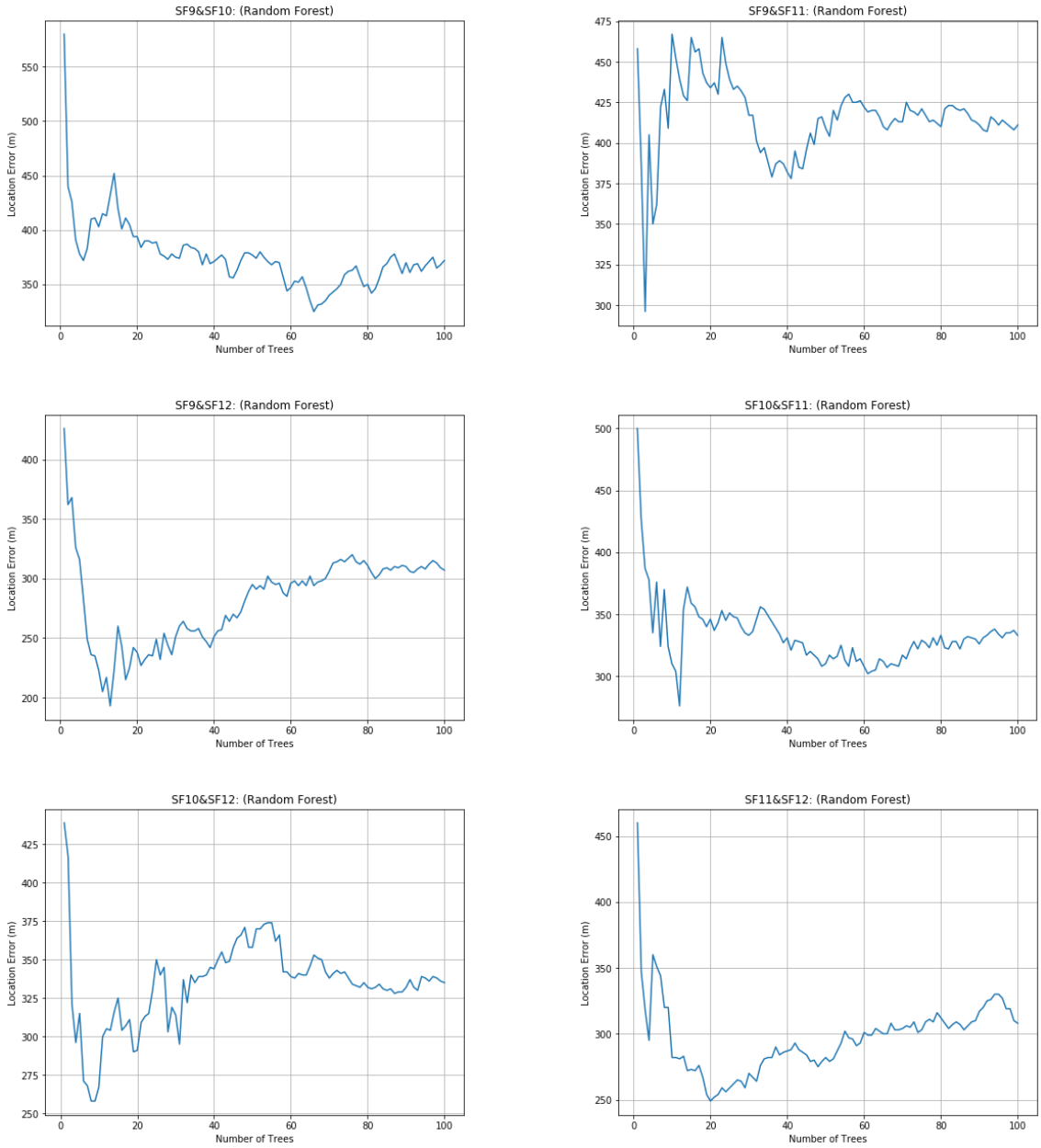



**Figure 6.3: Optimal number of trees in RF for Combined data (2 SFs).**

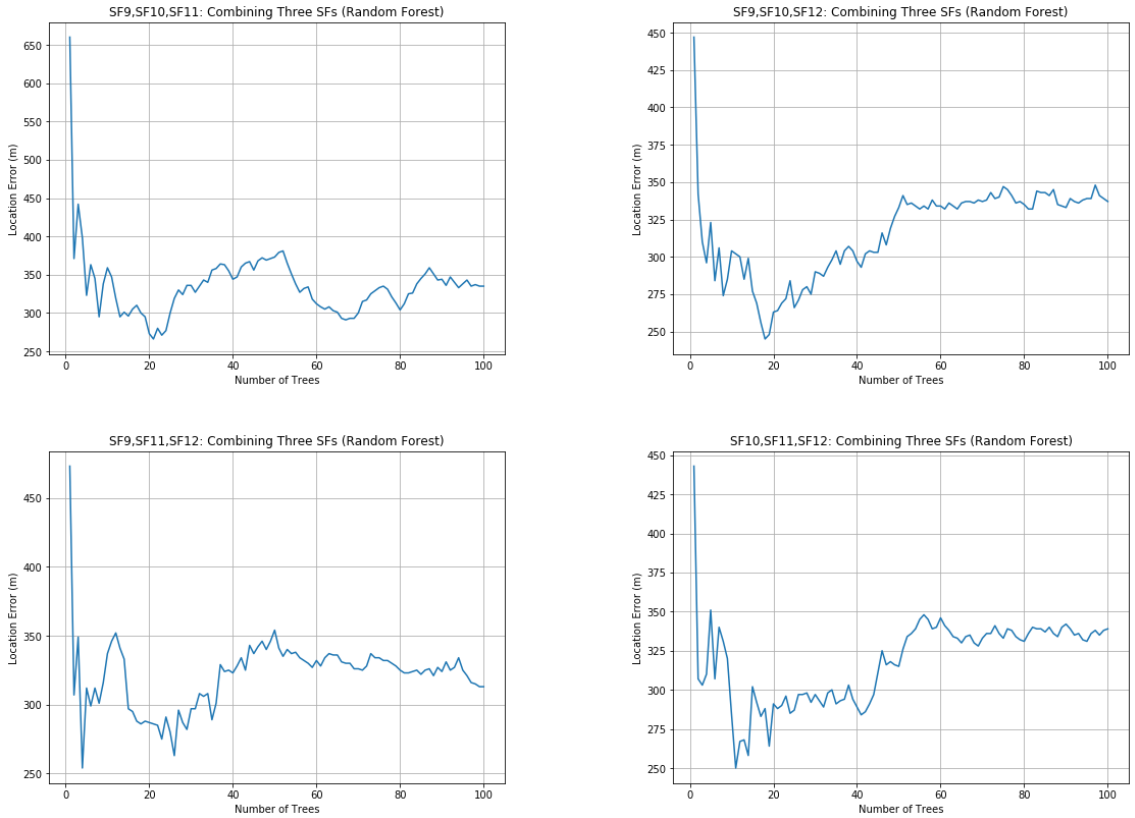



**Figure 6.4: Optimal number of trees in RF for Combined data (3 SFs).**

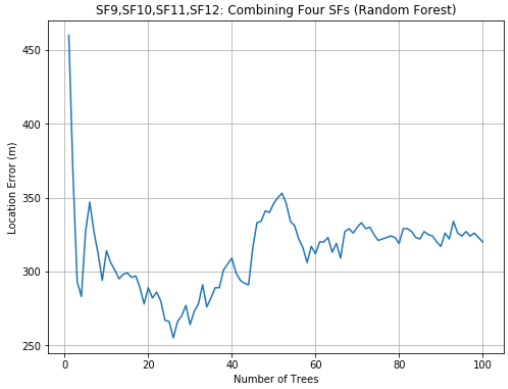



**Figure 6.5: Optimal number of trees in RF for Combined data (4 SFs).**

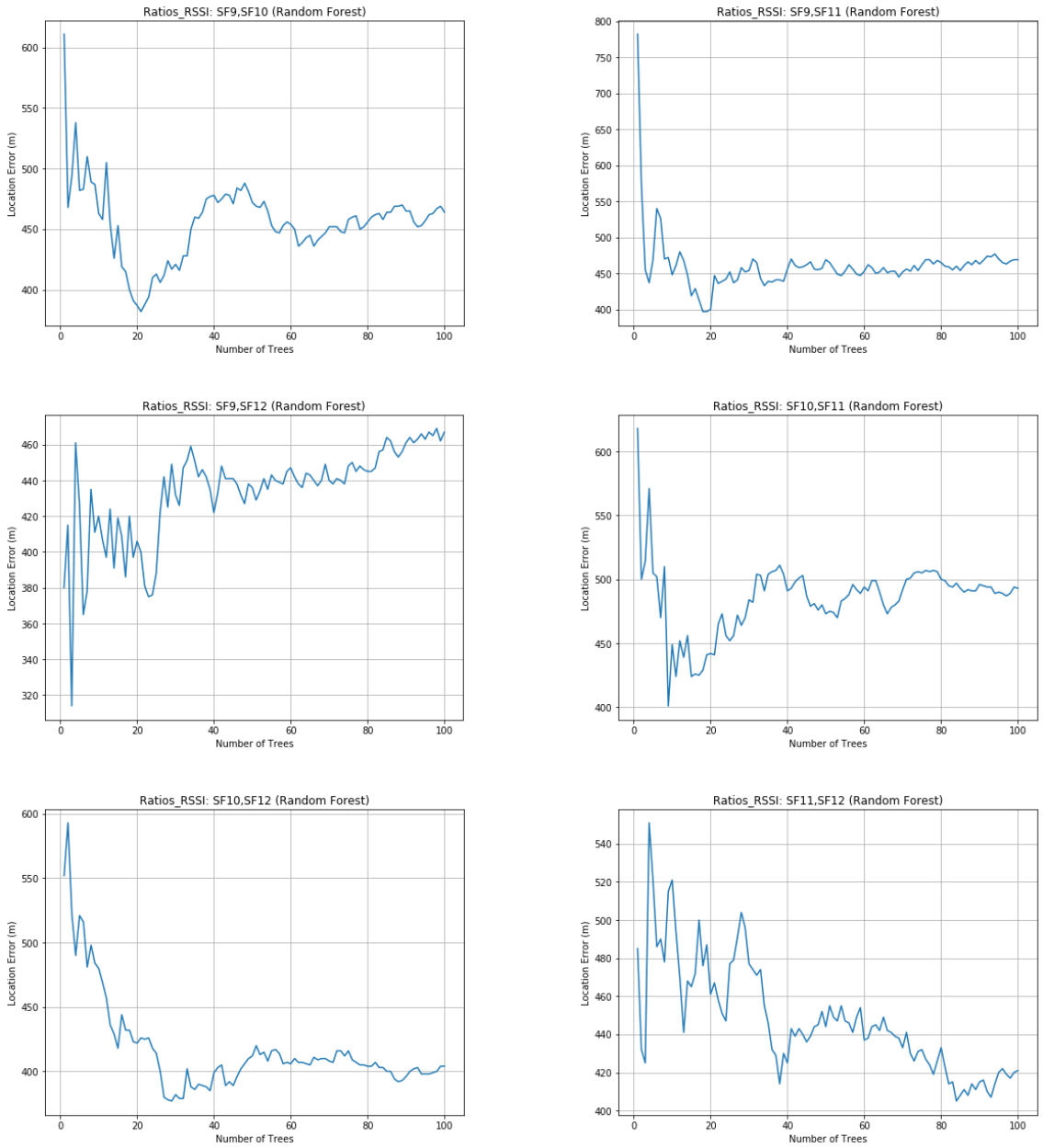



**Figure 6.6: Optimal number of trees in RF for Combined RSSI Ratio data (2 SFs).**

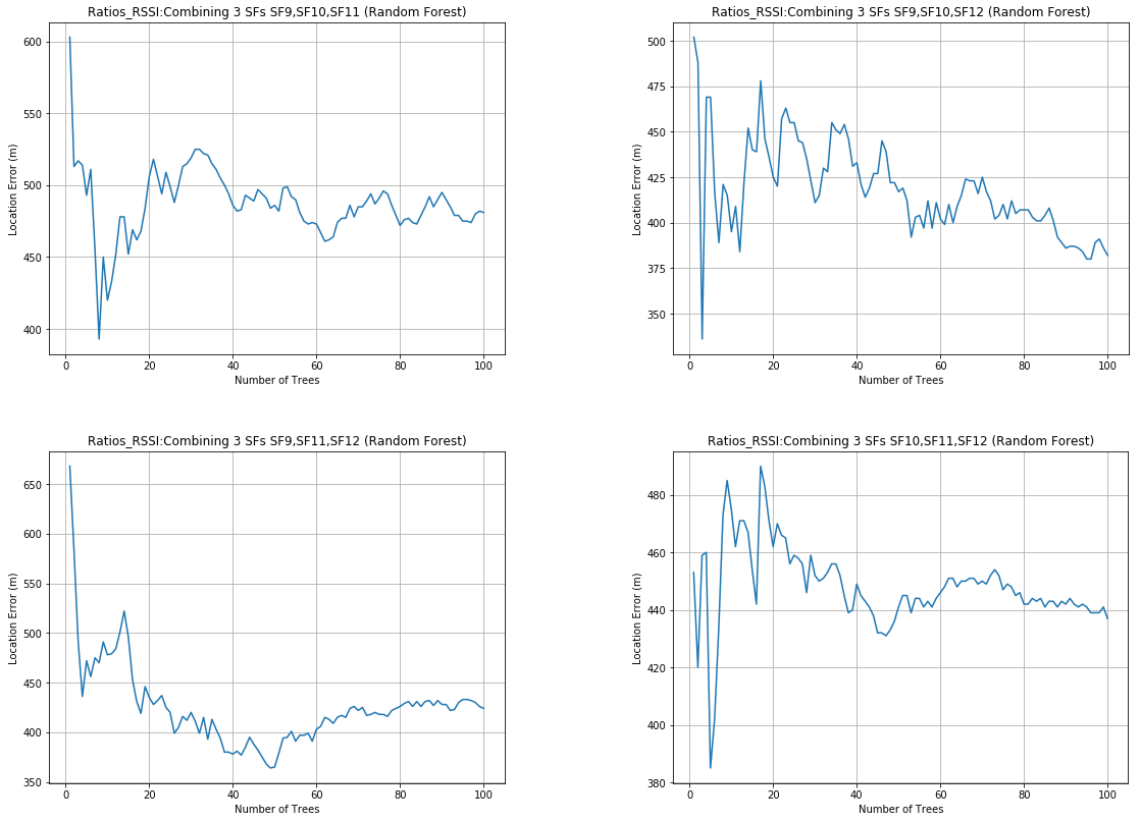



**Figure 6.7: Optimal number of trees in RF for Combined RSSI Ratio data (3 SFs).**

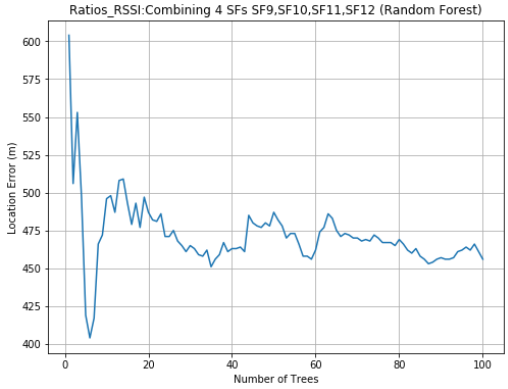



**Figure 6.8: Optimal number of trees in RF for Combined RSSI Ratio data (4 SFs).**

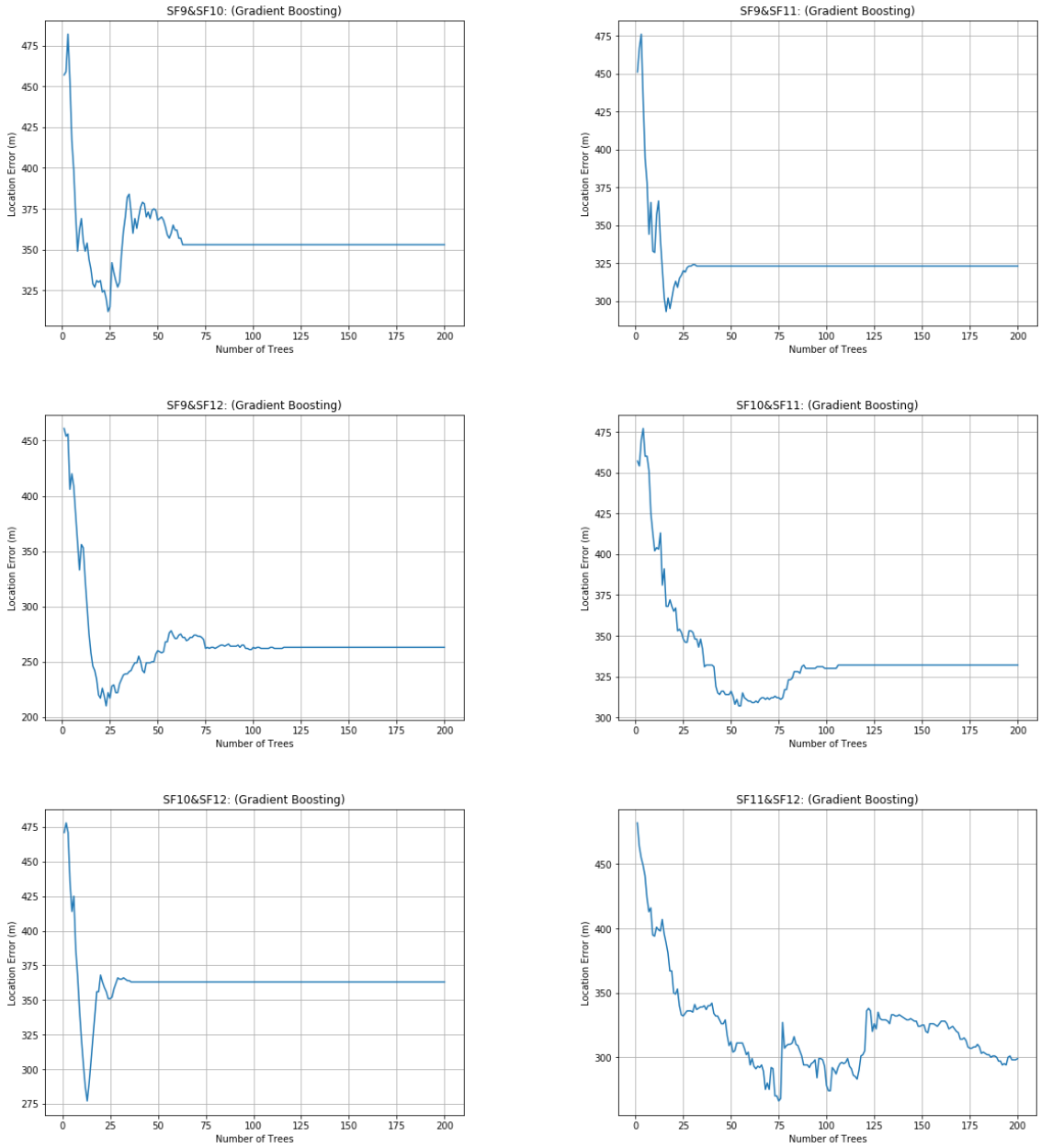



**Figure 6.9: Optimal number of trees in GB for Combined data (2 SFs).**

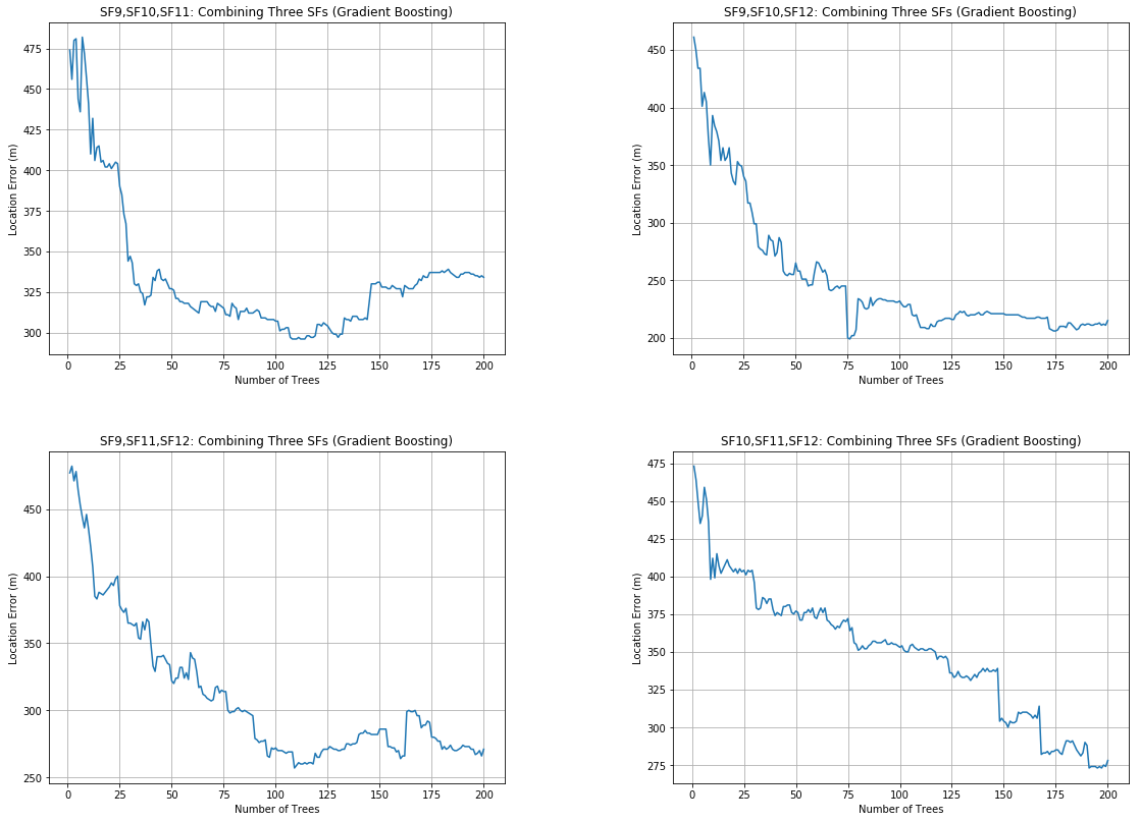



**Figure 6.10: Optimal number of trees in GB for Combined data (3 SFs).**

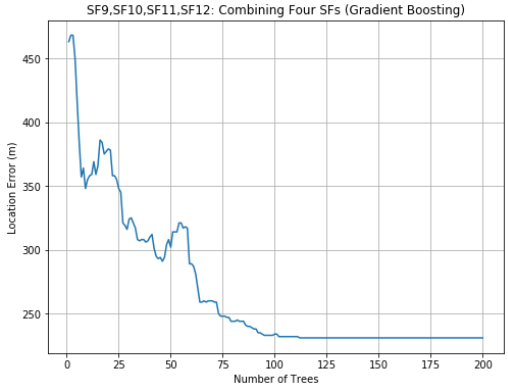



**Figure 6.11: Optimal number of trees in GB for Combined data (4 SFs).**

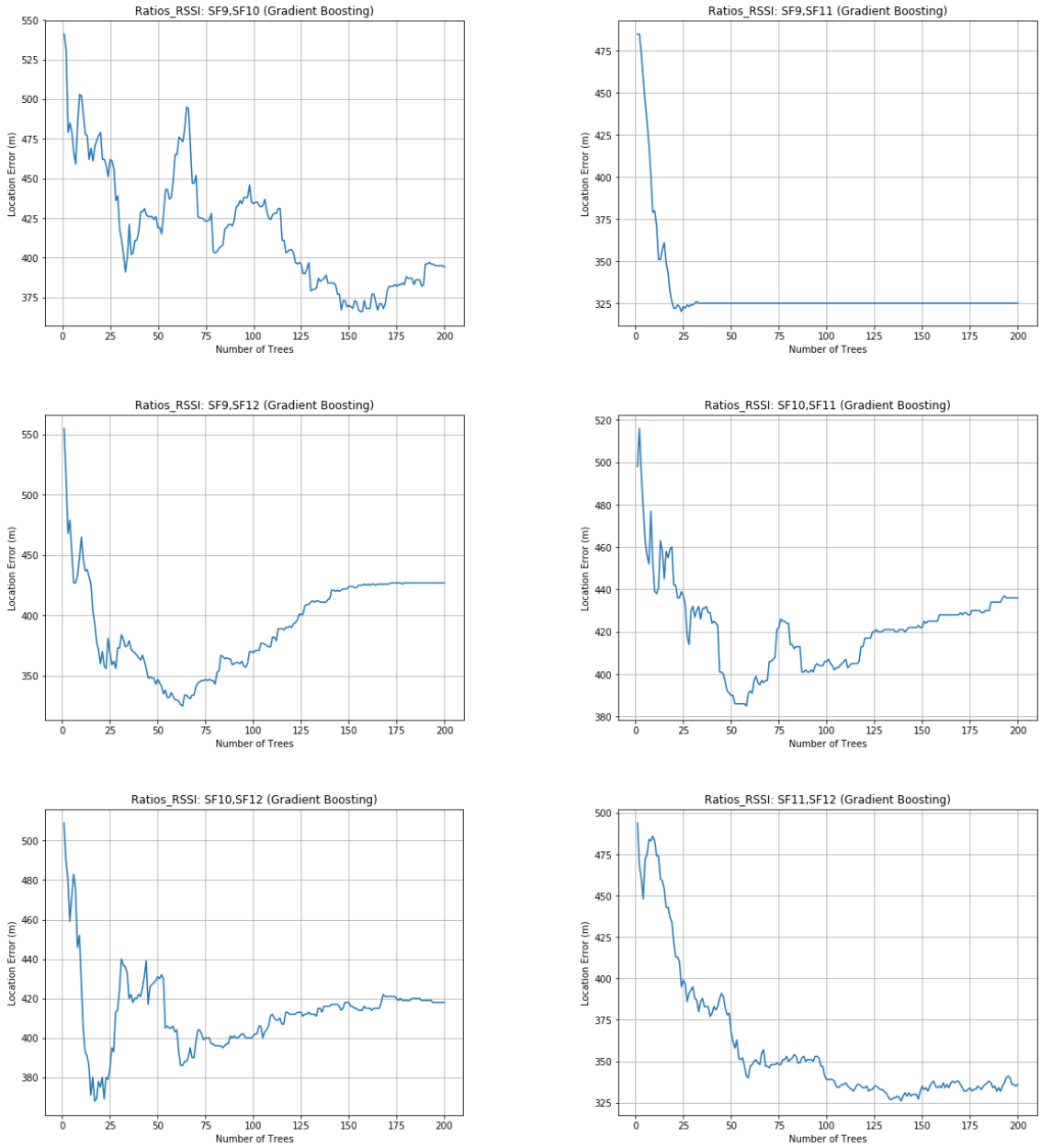



**Figure 6.12: Optimal number of trees in GB for Combined RSSI Ratio data (2 SFs).**

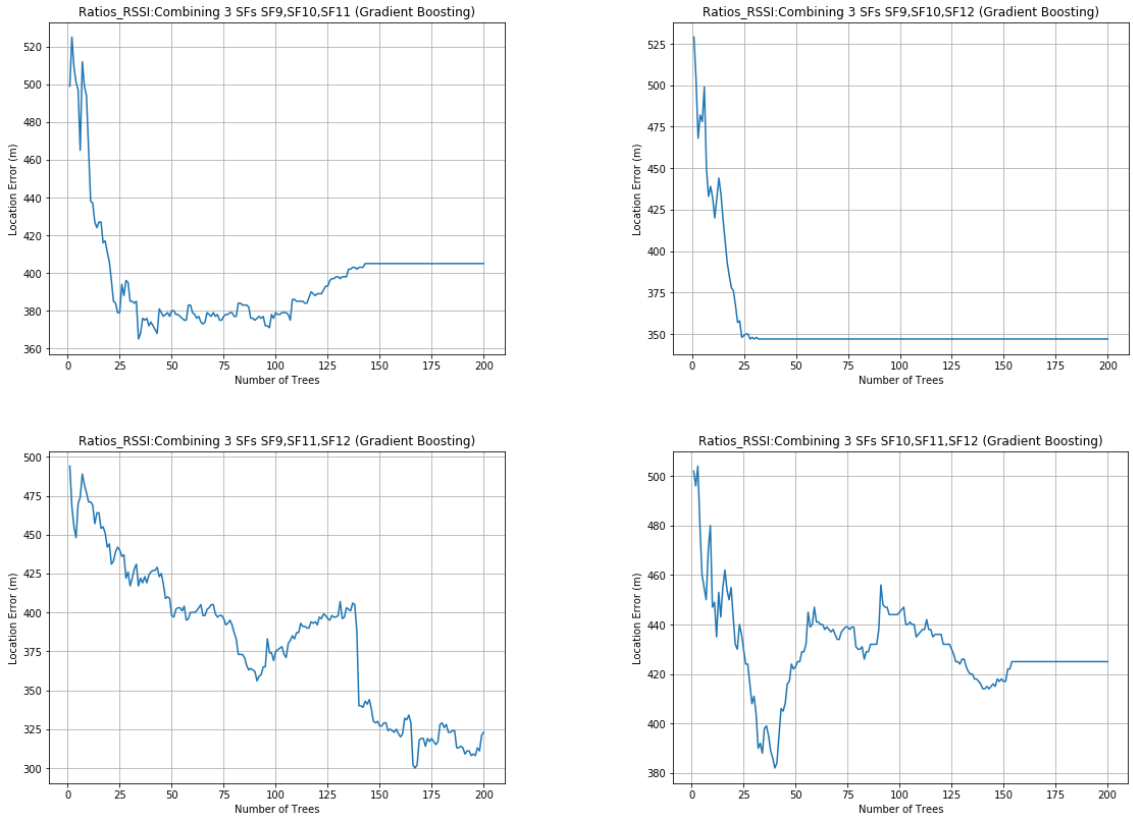



**Figure 6.13: Optimal number of trees in GB for Combined RSSI Ratio data (3 SFs).**

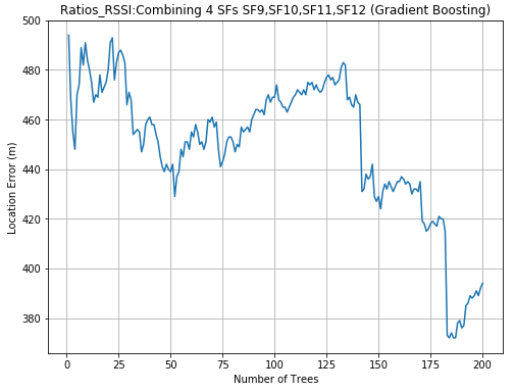



**Figure 6.14: Optimal number of trees in GB for Combined RSSI Ratio data (4 SFs).**

### *6.6.3 Impact of localisation algorithm*

A RDT approach has been used to build two ensemble learning models; RF and GB for node localisation. The results of the localisation performance have been compared with a base algorithm, regression tree model; here, only the combination of RSSI values by 2SFs is used to compare the algorithms.

The ensemble learning algorithms outperformed the regression decision tree for any combination of the RSSI features by the spreading factors; a single tree model is shown in Table 6.5. In terms of mean and median location error, RF has been shown to be more accurate with mean and median localisation error of 338m and 193m respectively when SF9 and SF12 features are combined, a 36.8% and 61% improvement in precision results compared with regression tree model. GB has the least minimum error of 22m when the features of SF9 and SF12 are combined. Both ensemble methods have consistent performance in terms of the median localisation error; for every feature combination, their median localisation error is less than 350m.

**Table 6.5: Performance of node localisation using combined RSSI features.**

| Models | SF9&SF10 (m) | SF9&SF11 (m) | SF9&SF12 (m) | SF10&SF11 (m) | SF10&SF12 (m) | SF11&SF12 (m) |
|---|---|---|---|---|---|---|
| DT(min) | 77 | 16 | 88 | 71 | 70 | 91 |
| DT(mean) | 555 | 530 | 535 | 487 | 486 | 546 |
| DT(median) | 477 | 416 | 495 | 368 | 398 | 470 |
| DT(RMSE) | 641 | 680 | 619 | 597 | 608 | 673 |
| RF(min) | 23 | 111 | 51 | 41 | 43 | 85 |
| RF(mean) | 383 | 448 | 338 | 396 | 329 | 354 |
| RF(median) | 325 | 296 | 193 | 276 | 258 | 249 |
| RF(RMSE) | 440 | 546 | 433 | 465 | 404 | 427 |
| GBR(min) | 18 | 54 | 22 | 46 | 119 | 70 |
| GBR(mean) | 397 | 437 | 344 | 395 | 401 | 356 |
| GBR(median) | 312 | 293 | 210 | 307 | 277 | 266 |
| GBR(RMSE) | 486 | 551 | 423 | 470 | 462 | 409 |

The CDF and the box plot of localisation error for DRT, RF and GB are shown in Figure 6.15, Figure 6.16, and Figure 6.17 respectively. The ensemble methods achieved the best performance yielding 75% localisation precision with error less than 454 m (RF)and 481 m (GBR). In contrast, the DRT achieved 75% localisation precision with error up to 600m.

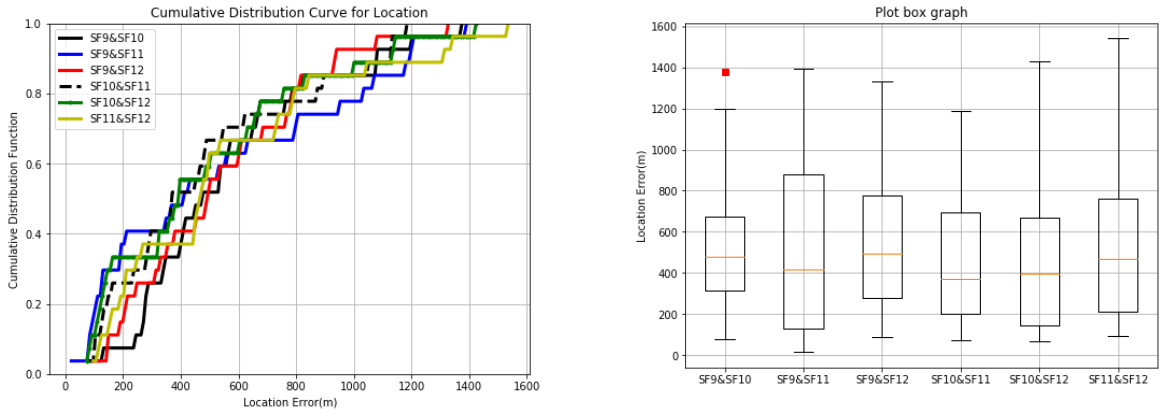



**Figure 6.15: CDF and box plot of localisation error for DRT using combined RSSI features.**

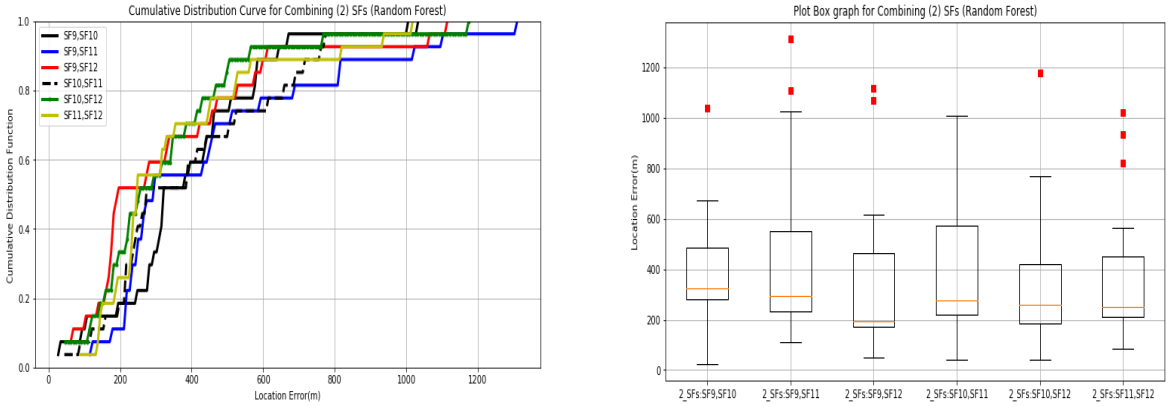



**Figure 6.16: CDF and box plot of localisation error for RF using combined RSSI features.**

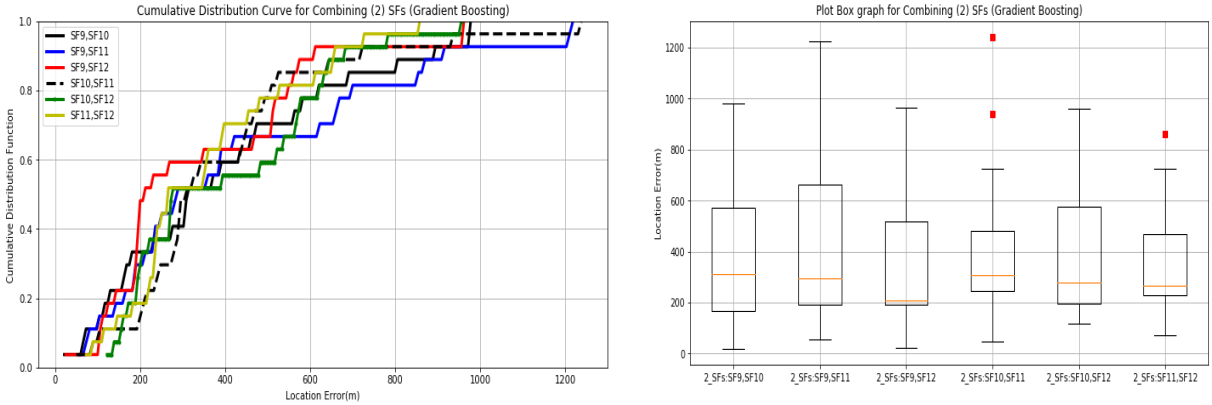



**Figure 6.17: CDF and box plot of localisation error for GB using combined RSSI features.**

### *6.6.4 Impact of Feature Combination*

In order to evaluate the impact of features on the performance of the node localisation models, an empirical evaluation of the regression decision tree and the ensemble models as a function of RSSI features of each spreading factor is carried out and compared to the performance on a combination of the RSSI features by spreading factors. In addition, the RSSI ratios used in Chapter 5 have been combined by spreading factors and used to investigate the consistency of the proposed feature transformation. The feature combination technique by spreading factor is extended to combining three and then four spreading SF to further investigate their impact. In order to have an unbiased evaluation, the ensemble methods are first evaluated using RSSI values at a single spreading factor. Table 6.6 shows the statistical performance of the models for single spreading factor data.

**Table 6.6: Performance of node localisation using single spreading factor RSSI features.**

| Models | SF9 (m) | SF10 (m) | SF11 (m) | SF12 (m) |
|---|---|---|---|---|
| GBR (min) | 107 | 43 | 35 | 38 |
| GBR (mean) | 498 | 513 | 435 | 349 |
| GBR(median) | 467 | 498 | 343 | 257 |
| GBR (RMSE) | 551 | 594 | 538 | 429 |
| RFR (min) | 72 | 77 | 115 | 16 |
| RFR (mean) | 464 | 486 | 465 | 370 |
| RFR (median) | 455 | 431 | 365 | 263 |
| RFR (RMSE) | 533 | 568 | 552 | 472 |

As shown in Table 6.6 the best median localisation error using RSSI features at a single spreading factor SF9 through S12 is 257 m. In contrast, RF and GB with combined features (2SFs) provide the best median localisation error of 193 m and 210m respectively as shown in Table 6.5, enhancing precision by 24.9% and 18.3% over using RSSI at a single spreading factor.

For RSSI ratio feature combination, only the ensemble methods are implemented. Table 6.7 shows the statistical performance of the models on the combined ratio data for 2SFs. From Table 6.5 and Table 6.6, it is evident that combining RSSI values yielded an improvement in terms of localisation accuracy and robustness. RF and GB estimate node location with distance error of 193 m and 210 m or less respectively for 50% of the time; when RSSI ratio

features are combined, the least median localisation error for RF and GB are 314 m and 320 m respectively.

**Table 6.7: Performance of node localisation using combined RSSI Ratio features (2SFs).**

| Models (Ratios) | SF9&SF10 (m) | SF9&SF11 (m) | SF9&SF12 (m) | SF10&SF11 (m) | SF10&SF12 (m) | SF11&SF12 (m) |
|---|---|---|---|---|---|---|
| **RF(min)** | 20 | 39 | 12 | 60 | 20 | 20 |
| **RF(mean)** | 496 | 470 | 455 | 504 | 435 | 460 |
| **RF(median)** | 382 | 397 | 314 | 401 | 377 | 405 |
| **RF(RMSE)** | 586 | 548 | 566 | 595 | 492 | 543 |
| **GBR(min)** | 87 | 70 | 18 | 76 | 64 | 31 |
| **GBR(mean)** | 520 | 475 | 453 | 527 | 455 | 466 |
| **GBR(median)** | 366 | 320 | 325 | 385 | 368 | 326 |
| **GBR(RMSE)** | 605 | 596 | 526 | 608 | 509 | 580 |

The impact of feature combination was investigated further by combining 3 and 4 different SFs data, the results of which are summarised in Table 6.8, Table 6.9, and Table 6.10. It is clear that no significant improvement results when 3SFs or 4SFs data are combined compared to combining 2SFs.

**Table 6.8: Performance of node localisation using combined RSSI features (3SFs).**

| Models | SF9&SF10&SF11 (m) | SF9& SF10&SF12 (m) | SF9& SF11&SF12 (m) | SF10&SF11&SF12 (m) |
|---|---|---|---|---|
| **RF(min)** | 31 | 63 | 91 | 89 |
| **RF(mean)** | 368 | 334 | 384 | 319 |
| **RF(median)** | 266 | 245 | 254 | 250 |
| **RF(RMSE)** | 437 | 420 | 470 | 375 |
| **GBR(min)** | 30 | 28 | 89 | 60 |
| **GBR(mean)** | 371 | 330 | 335 | 352 |
| **GBR(median)** | 296 | 199 | 257 | 273 |
| **GBR(RMSE)** | 436 | 427 | 390 | 417 |

**Table 6.9: Performance of node localisation using combined RSSI Ratio features (3SFs).**

| Models | SF9&SF10&SF11 (m) | SF9& SF10&SF12 (m) | SF9& SF11&SF12 (m) | SF10&SF11&SF12 (m) |
|---|---|---|---|---|
| **RF(min)** | 63 | 111 | 47 | 64 |
| **RF(mean)** | 512 | 442 | 449 | 480 |
| **RF(median)** | 393 | 336 | 364 | 385 |
| **RF(RMSE)** | 593 | 543 | 530 | 563 |
| **GBR(min)** | 106 | 28 | 99 | 72 |
| **GBR(mean)** | 487 | 488 | 451 | 505 |
| **GBR(median)** | 365 | 347 | 300 | 382 |
| **GBR(RMSE)** | 577 | 603 | 560 | 593 |

**Table 6.10: Performance of node localisation using combined RSSI & RSSI Ratio features (4SFs).**

| Models | SF9&SF10&SF11&SF12 RSSI Values (m) | SF9& SF10&SF11&SF12 RSSI Ratios (m) |
|---|---|---|
| **RF(min)** | 58 | 28 |
| **RF(mean)** | 324 | 532 |
| **RF(median)** | 255 | 404 |
| **RF(RMSE)** | 396 | 617 |
| **GBR(min)** | 76 | 35 |
| **GBR(mean)** | 353 | 469 |
| **GBR(median)** | 231 | 372 |
| **GBR(RMSE)** | 435 | 550 |

Based on the results of the evaluation, it can be concluded that combination of features by spreading factor with ensemble learning methods can provide a higher localisation accuracy consistently on comparison with single spreading factor RSSI features. More specifically, combining data from 2SFs proved to yield the maximum improvement.

## 6.7 Summary

An investigation into the use of ensemble learning methods and combination of features of different spreading factors to improve the accuracy of node localisation has been carried out.

The spreading factor plays an important role on the RSSI received. Therefore, a combination of RSSI features from any two spreading factor can have a significant impact on the quality

of the features, providing richer data for node localisation thereby improving the accuracy. The performance of the combination of features on localisation has been evaluated. A combination of three and four spreading factor-RSSI features has been investigated and the results compared with using a combination of two spreading factor-RSSI features.

Furthermore, machine learning ensemble methods - Random Forest and Gradient Boosting Regression - have been used for localisation using the combined features. The use of combination of spreading factor-RSSI features in conjunction with ensemble learning techniques has proven to improve localisation performance in the selected environment. RF and GBR with combined features (2SFs) provide the best median localisation error of 193 m and 210m respectively, enhancing precision by 24.9% and 18.3% over using RSSI features at a single spreading factor with best median error of 257 m. RF has been shown to be more accurate when SF9 and SF12 features are combined. RF provided a 61% improvement in precision results compared with results from the regression tree model.

Compared with KNN and WKNN, which rely on the storage all of the data in the fingerprint database and piecemeal one-by-one comparisons in the online matching stage, RF and GB methods only need to keep the trained model and carry out the node splitting process in the online matching phase.

# CHAPTER 7 CONCLUSIONS AND FUTURE WORK

## 7.1 Conclusions

The Thesis presented the development and performance evaluation of a cost-effective IoT-implemented node localisation technique for suburban environments subject to sandstorm conditions in Saudi Arabia. Results demonstrate the attainable performance of RSSI-based node localisation in dynamic environments through a limited number of Gateways and significant distances between nodes and Gateways.

The network-based approach utilised a LoRaWAN implementation as the target application required coverage over a region characterised by significant distances between nodes. In order to overcome the challenges imposed by the operational environment and limited network resources, the solution selected the readily attainable Received Signal Strength Indicator (RSSI) as the basis for deriving node locations. Furthermore, localisation was implemented through fingerprinting and its performance enhanced through machine learning techniques.

The evolution of the development was informed throughout by results obtained starting from the foundation data gathering phase and the subsequent results on the enhancements owing to the application of a number of machine learning techniques with RSSI-derived feature inputs to establish the fingerprints. A key early output was the validation of the importance of key LoRaWAN parameters, most notably the Spreading Factor, to the performance of node localisation. Different strategies such as ensemble methods and feature

engineering were evaluated for enhanced estimation. Table 7.1 summarises the optimal node localisation performance for the developed techniques.

**Table 7.1: Summary table of the optimal performance results for the developed models.**

| | Chapter | Model | RSSI feature | Optimal Spreading Factor | Optimal Performance (Median) |
|---|---|---|---|---|---|
| **1** | Chapter 4 | KNN | Single RSSI | SF12 | 316 m |
| **2** | Chapter 4 | WKNN | Single RSSI | SF12 | 315 m |
| **3** | Chapter 5 | SVR | Ratio RSSI | SF11 | 303 m |
| **4** | Chapter 5 | GPR | Ratio RSSI | SF12 | 317 m |
| **5** | Chapter 6 | GBR | Combined RSSI | SF9 and SF12 | 210 m |
| **6** | Chapter 6 | RF | Combined RSSI | SF9 and SF12 | 193 m |

Chapter Four details the implementation and performance evaluation of RSSI-based location fingerprinting. RSSI is considered as the most readily available in comparison with other transmitted parameter, in the goal of creating unique signatures for each node location. Confirmation that basic path transmission models, such as the classical two-ray propagation, yield poor estimates of distance stimulated to use of machine learning to improve model accuracy. K-nearest neighbour algorithm and its variant, weighted K-nearest neighbour were used as baseline machine learning algorithms to develop localisation models; both these models demonstrated an improvement over established propagation models.

KNN and WkNN models yielded improvement in localisation accuracy as the SF increased from SF9 to SF12. Further the trade-off between latency and accuracy at the highest SF12 is argued to be desirable in the application under consideration. Shadowing and reflections that impair consistent reception is more likely at lower than higher SFs. The median optimal performance of KNN and WKNN were 316m and 315m respectively at SF12 (Table 7.1). In this respect, the main contributions are as follows;

- implementation of fingerprinting technique for IoT node localisation in suburban environment subject to sandstorms using RSSI as inputs to the fingerprints enhanced with KNN algorithms. The approach is general and can be applied to any localisation scenario.
- characterisation of the impact of LoRa SF on the reception performance of LoRaWAN nodes in a challenging environment.

Chapter 5 details node localisation performance based on the use of relative RSSIs as inputs to fingerprints in tandem with more advanced machine learning techniques. The approach targets a reduction in the temporal variations in RSSIs e.g. owing to multi-paths. The foundation RSSI values are transformed into RSSI ratios by pairs of Gateways. Results show that RSSI ratios improve performance (Table 7.1). Two kernel-based algorithms - Support Vector Regression (SVR) and Gaussian Process Regression (GPR) - are parameterised to model the relationship between RSSI ratios and reference node location. Moreover, analyses to determine the impact of kernel functions demonstrated that new functions derived from a combination of existing kernels yield more accurate estimations compared to existing

kernels. SVR has the lowest median error of 303m compare to 317m for GPR at SF11 and SF12 respectively (Table 7.1). The main contributions of the chapter are as follows:

- use of relative RSSI as location fingerprints in conjunction with advanced machine learning techniques for improved node localisation
- effect of combined kernel functions on the performance of node localisation

Chapter 6 presents two new techniques based on feature engineering and ensemble methods to improve the fingerprint; both methods demonstrated improvement over preceding techniques. RSSI at different SFs are combined to form new fingerprints. The combination of two different parameters - RSSI and RSSI ratio – is tested to determine the impact on the quality of the features in the goal to establish a richer database for localisation. Results corroborate that combined features are more robust and enhance performance. Furthermore, a combination of three and four spreading factor-RSSI features were explored in order to understand the combination strategy that best preserves a richer location information. However a combination of more than two different spreading factor-RSSI introduces noise into the fingerprints and hence compromises the performance of the node localisation system.

The second strategy was the use of a machine learning ensemble technique. Two tree-based ensemble methods - Gradient Boosting Regression and Random Forest were identified and used to model the complex relationship between the combined RSSI fingerprints and the reference node locations. The evaluation of the machine learning ensemble methods demonstrated manifest improvement in node localisation accuracy compared with using

single tree regression model. RF has been shown to be more accurate with median localisation error of 193m when SF9 and SF12 features are combined, a 61% improvement in precision results compared with results from the regression tree model (Table 7.1). The combination of machine learning and spreading factor-RSSI combined features in the field of node localisation has not been reported extensively to date. The main contributions within the Chapter are as follows:

- a novel feature engineering strategy is introduced for enhanced node localisation in challenging propagation environments. The feature combination scheme presented is modular and can be used in other applications and environments.
- development of two machine learning ensemble models through a limited number of Gateways and significant distances between nodes and Gateways.

Evidence in support of proving the feasibility of providing an energy-efficient localisation technique with acceptable performance for extensive IoT applications and services within challenging operational environments is provided. Feature engineering methods based on the readily attainable Received Signal Strength Indicator (RSSI) acquired within a LoRaWAN network setting provide a spectrum of options for the estimation of node location that form the basis for the development of solutions addressing multiple applications operating in many diverse environments.

## 7.2 Future Work

The knowledge gained has surfaced a number of additional research challenges;

- the proposed localisation solution considered a scenario of fixed nodes; future work should address the challenges associated with mobile node localisation
- consideration of the impact of other LoRa parameters on the performance of node localisation extending the analysis beyond SF only
- only the RSSI of the received packets are extracted and used as location fingerprints; future work could consider other parameters.
- ML has been applied in the localisation solution to primarily optimise accuracy. Future development should be orientated toward creating node localisation models focused on the joint optimisation of both accuracy and latency
- an aspect not considered in the present research is the complexity of the algorithms, important if the numbers of nodes and Gateways increase

# REFERENCES


[1] C. Chang, S. Srirama and R. Buyya, "Internet of Things (IoT) and new computing paradigms," *Fog and Edge Computing: Principles and Paradigms,* pp. 1-23, 2019.

[2] M. James, C. Michael, B. Jacques, D. Richard, B. Peter, M. Alex, W. Jonathan and A. Dan, "THE INTERNET OF THINGS: MAPPING THE VALUE," *USA: McKinsey Global Instit,* vol. San Francisco, no. CA, 2015.

[3] K. Briony, K. ANN-KATHRIN, P.-M. AMNA, E. JOHN, S. PETAR and C. JAMES, "Statista:Internet of Things - number of connected devices worldwide 2015-2025," 27 November 2016. [Online]. Available: https://www.statista.com/statistics/471264/iot-number-of-connected-devices-worldwide/. [Accessed 2019].

[4] USDoD, "Global Positioning System Standard Positioning Service Performance Standard," United States Department of Defense, Washington DC, USA, 2008.

[5] T. Ahmad, X. J. Li and B. C. Seet, "Parametric loop division for 3D localization in wireless sensor networks," *sensors Article,* vol. 17, no. 7, pp. 1-32, 2017.

[6] T. He, C. Huang, B. Blum, J. Stankovic and T. Abdelzaher, "Range-free localization schemes for large scale sensor networks," San Diego, CA, USA, 2003.

[7] C. Laoudias, A. Moreira, S. Kim, S. Lee, L. Wirola and C. Fischione, "A Survey of Enabling Technologies for Network Localization, Tracking, and Navigation," *IEEE Communications Surveys & Tutorials,* vol. 20, no. 4, pp. 3607 - 3644, 2018.

[8] C. Frank, S. Krishnamurthy, K. Stewart and X. a. L. R. Zhuang, "Method and apparatus for generating reference signals for accurate time-difference of arrival estimation," *Google Technology Holdings LLC,* no. U.S. Patent 9,541,632, 2017.

[9] R. Peng and S. M. L., "Angle of Arrival Localization for Wireless Sensor Networks," *IEEE Communications Society on Sensor and Ad Hoc Communications and Network,* vol. 1, pp. 374-382, 2006.

[10] S. Sadowski and P. Spachos, "Rssi-based indoor localization with the internet of things," *IEEE Access,* vol. 6, pp. 30149-30161, 2018.

[11] A. Mdhaffar, T. Chaari, K. Larbi, M. Jmaiel and B. Freisleben, "IoT-based health monitoring via LoRaWAN," Ohrid, Macedonia,, 2017.

[12] J. Wan, M. Al-awlaqi, M. Li, M. O'Grady, X. Gu, J. Wang and N. Cao, "Wearable IoT enabled real-time health monitoring system," *EURASIP J. Wirel. Commun. Netw.,* vol. 1, no. 298, 2018.

[13] M. Zamora-Izquierdo, J. Santa, J. Martínez, V. Martínez and A. Skarmeta, "Smart farming IoT platform based on edge and cloud computing.," *Biosyst. Eng.,* vol. 177, pp. 4-17, 2019.

[14] M. Tanaka, Y. Miyanishi, M. Toyota, T. Murakami, R. Hirazakura and T. Itou, “A study of bus location system using LoRa: Bus location system for community bus ‘notty’.,” Nagoya, Japan, 2017.

[15] J. James and S. Nair, “Efficient, real-time tracking of public transport, using LoRaWAN and RF transceivers,” Penang, Malaysia, 2017.

[16] C. Wongeun, C. Yoon-Seop, J. Yeonuk and S. Junkeun, “Low-Power LoRa Signal-Based Outdoor Positioning Using Fingerprint Algorithm,” *SPRS International Journal of Geo-Information,* vol. 7, no. 11, pp. 1-15, 2018.

[17] R. S. Sinha, Y. Wei and S.-H. Hwang, “A survey on LPWA technology: LoRa and NB-IoT,” *Elsevier B.V,* pp. 14-21, 2017.

[18] U. Raza, P. Kulkarni and M. Sooriyabandara, “Low Power Wide Area Networks: An Overview,” *IEEE Communications Surveys & Tutorials,* pp. 855-873, 2017.

[19] BehrTech, “6 Leading Types of IoT Wireless Tech and Their Best Use Cases,” 2020. [Online]. Available: https://behrtech.com/blog/6-leading-types-of-iot-wireless-tech-and-their-best-use-cases/. [Accessed 22 2 2020].

[20] R. O. Andrade and S. G. Yoo, “A Comprehensive Study of the Use of LoRa in the Development of Smart Cities,” *Appled Sciences,* vol. 9, no. 4753, pp. 1-39, 2019.

[21] K. Mekki, E. Bajic, F. Chaxel and F. Meyer, “Overview of Cellular LPWAN Technologies for IoT Deployment: Sigfox, LoRaWAN, and NB-IoT,” Athens, 2018.

[22] M. Aernouts, R. Berkvens and K. &. W. M. Van Vlaenderen, “Sigfox and LoRaWAN Datasets for Fingerprint Localization in Large Urban and Rural Areas,” no. Data, p. 13, 2018.

[23] M. Lauridsen, B. Vejlgaard, I. Z. Kovacs, H. Nguyen and P. Mogensen, “Interference measurements in the European 868 MHz ISM band with focus on LoRa and SigFox,” San Francisco, CA, 2017.

[24] E. Migabo, K. Djouani, A. Kurien and T. Olwal, “A Comparative Survey Study on LPWA Networks: LoRa and NB–IoT,” *ICT Express,* vol. 3, no. 1, pp. 14-21, 2017.

[25] K. Mekkia, E. Bajica, F. Chaxela and F. Meyerb, “A comparative study of LPWAN technologies for large-scale IoT,” *ICT Express ,* vol. 5, no. 1, pp. 1-7, 2019.

[26] J. Haxhibeqiri, E. Poorter, I. Moerman and J. Hoebeke, “A Survey of LoRaWAN for IoT: From Technology to Application,” *Sensors,* vol. 18, no. 11, 2018.

[27] J. Paredes-Parra, A. García-Sánchez, A. Mateo-Aroca and A. Molina-Garcia, “An Alternative Internet-of-Things Solution Based on LoRa for PV Power Plants: Data Monitoring and Management.,” *Energies,* vol. 12, no. 881, 2019.

[28] F. Adelantado, X. Vilajosana, P. Tuset-Peiro, B. Martinez and J. Melia, "Understanding the limits of LoRaWAN," *IEEE Communications Magazine,* pp. 1-5, 2017.

[29] semtech, "semtech," 2019. [Online]. Available: https://www.semtech.com/uploads/images/LoRa_Why_Range.png. [Accessed 25 12 2019].

[30] T. M. Workgroup, "LoRaWAN™ What is it? A technical overview of LoRa® and LoRaWAN™," November 2015. [Online]. Available: https://lora-alliance.org/sites/default/files/2018-04/what-is-lorawan.pdf. [Accessed February 2020].

[31] M. Centenaro, L. Vangelista, A. Zanella and M. Zorzi, "Long-range communications in unlicensed bands: the rising stars in the IoT and smart city scenarios," *IEEE Wireless Communications,* vol. 23, no. 5, pp. 60-67, 2016.

[32] W. Bakkali, M. Kieffer, M. Lalam and T. Lestable, "Kalman filter-based localization for Internet of Things LoRaWAN™ end points," in *IEEE 28th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*, Montreal, QC, 2017.

[33] M. Muzammir, H. Abidin, S. Abdullah and F. Zaman, "Performance Analysis of LoRaWAN for Indoor Application," Malaysia, 2019.

[34] K. Staniec and M. Kowal, " LoRa Performance under Variable Interference and Heavy-Multipath Conditions," *Wireless Communications and Mobile Computing,* vol. 2018, pp. 1-9, 2018.

[35] B. Islam, M. Islam and S. Nirjon, "Feasibility of LoRa for Indoor Localization," 2017. [Online]. Available: https: //pdfs.semanticscholar.org/ab00/c1eacbdd76732b7438ec8e5653f7c875def4.pdf. [Accessed 22 July 2019].

[36] L. Nordin, "What is LoRa and LoRaWAN?," [Online]. Available: https://zakelijkforum.kpn.com/lora-forum-16/what-is-lora-and-lorawan-8314. [Accessed 23 1 2020].

[37] A. Lavric and V. Popa, "Internet of Things and LoRa™ Low-Power Wide-Area Networks: A survey," Iasi, 2017.

[38] s. soft, "www.simplesoft.com," 2019. [Online]. Available: https://www.simplesoft.com/SimpleIoTSimulatorForLoraWan.html . [Accessed 17 July 2019].

[39] S. Tarkoma and A. Katasonov, "Internet of Things Strategic Research Agenda (IoT-SRA)," *Finnish Strategic Centre for Science, Technology, and Innovation: For Information and Communications (ICT) Services, businesses, and technologies,* 2011.

[40] A. Springer, W. Gugler, M. Huemer, L. Reind, C. Ruppel and R. Weigel, "Spread spectrum communications using chirp signals," *Proceedings of the IEEE/AFCEA Information Systems for Enhanced Public Safety and Security (EUROCOMM 2000), and Security (EUROCOMM 2000), Munich, Germany,* pp. 166-170, 19 May 2000.

[41] A. Augustin, J. Yi, T. Clausen and W. M. Townsley, "A Study of LoRa: Long Range & Low Power Networks for the Internet of Things," *Sensors,* vol. 16, no. 9, p. 1466, 2016.

[42] C. Raaltenpark, "Chirp spread spectrum," 2 October 2018. [Online]. Available: https://www.mobilefish.com/download/lora/lora_part12.pdf. [Accessed 21 February 2020].

[43] J. Petäjäjärvi, K. Mikhaylov, M. Pettissalo, J. Janhunen and J. Iinatti, "Performance of a low-power wide-area network based on LoRa technology: Doppler robustness, scalability, and coverage," *International Journal of Distributed Sensor Networks,* vol. 13, no. 3, 2017.

[44] S. Ghoslya, "All About LoRa and LoRaWAN," 2017. [Online]. Available: http://www.sghoslya.com/p/lora-is-chirp-spread-spectrum.html. [Accessed 7 October 2018].

[45] Semtech Corporation, "LoRa SX1276/77/78/79 Datasheet, Rev. 5," August 2016. [Online]. Available: http://www.semtech.com/images/datasheet/sx1276.pdf. [Accessed 17 May 2017].

[46] Semtech Corporation, “What is LoRa?,” 2017. [Online]. Available: http://www.semtech.com/wireless-rf/internet-of-things/what-is-lora/. [Accessed 17 May 2017].

[47] LoRa Alliance, “LoRa Specification,” LoRa Alliance, Inc, San Ramon, CA, USA, 2015.

[48] C. Fehri, M. Kassab, S. Abdellatif, P. Berthou and A. Belghith, “LoRa technology MAC layer operations and Research issues,” *Procedia Computer Science,* vol. 130, pp. 1096-1101, 2018.

[49] E. D. Kaplan and H. C. J., Understanding GPS: Principles and Applications, 2nd ed., Norwood, MA, USA: Artech House, 2005.

[50] H.-W. Bernhard, L. Herbert and W. Elmar, GNSS – Global Navigation Satellite Systems,GPS, GLONASS, Galileo, and more, 1 ed., Springer-Verlag Wien, 2008.

[51] GPS.GOV, “ Official U.S. government information about the Global Positioning System (GPS) and related topics,” 2019. [Online]. Available: https://www.gps.gov/systems/gps/space/. [Accessed 14 July 2019].

[52] A. Zaidi and M. Suddle, “Global Navigation Satellite Systems: A Survey,” Pakistan, 2006.

[53] C. Ning, R. Li and L. Kejiong, "Outdoor Location Estimation Using Received Signal Strength-Based Fingerprinting," *Wireless Personal Communications,* vol. 89, p. 365–384, 2016.

[54] S. Gopi, Global Positioning System: Principles And Applications, New Delhi: Tata McGraw-Hill Education, 2005, p. 163.

[55] A. Bensky, Wireless positioning technologies and applications, Norwood, MA 02062: Artech House, 2008.

[56] B. Fargas and M. Petersen, "GPS-free geolocation using LoRa in low-power WANs," *Global Internet of Things Summit (GIoTS),* no. IEEE, pp. 1-6, 2017.

[57] F. Diggelen, "A-GPS: Assisted GPS, GNSS, and SBAS (GNSS Technology and Applications)," 6th, Ed., Norwood, MA, Artech House Publishers, 2009.

[58] G. Huang, D. Akopian and C. L. P. Chen, "Measurement and Modeling of Network Delays for MS-Based A-GPS Assistance Delivery," *IEEE TRANSACTIONS ON INSTRUMENTATION AND MEASUREMENT,* vol. 36, no. 8, pp. 1896-1906, 2014.

[59] A. Awad and S. Mohan, "Internet of Things for a Smart Transportation System," *International Journal of Interdisciplinary Telecommunications and Networking,* vol. 11, no. 1, pp. 57-70, 2019.

[60] R. M. Buehrer and S. Venkatesh, “Chapter 6: Fundamentals of Time-of-Arrival Position Location,” in *Handbook of Position Location: Theory, Practice, and Advances*, S. A. R. Zekavat and R. M. Buehrer, Eds., Piscataway, NJ, USA, IEEE, 2012, pp. 175-212.

[61] R. Henriksson, “Indoor positioning in LoRaWAN networks: Evaluation of RSS positioning in LoRaWAN networks using commercially available hardware,” Department of Signals and System, Chalmers University of Technology, MSc thesis, Göteborg, Sweden, 2016.

[62] N. Podevijn, D. Plets, J. Trogh, L. Martens, P. Suanet, K. Hendrikse and W. Joseph, “TDoA-based outdoor positioning with tracking algorithm in a public LoRa network,” London, 2018.

[63] S. C. LoRa Alliance™, “LoRaWAN GEOLOCATION WHITEPAPER,” January 2018. [Online]. Available: https://lora-alliance.org/sites/default/files/2018-04/geolocation_whitepaper.pdf. [Accessed 22 February 2020].

[64] C. Gu, L. Jiang and R. Tan, “LoRa-Based Localization: Opportunities and Challenges,” in *The 1st Workshop on Low Power Wide Area Networks for Internet of Things (LPNET)*, Beijing, 2019.

[65] T. Chan, F. Chan, W. Read, B. Jackson and B. Lee, “Hybrid localization of an emitter by combining angle-of-arrival and received signal strength measurements,” Toronto, Canada,, 2014.

[66] W. Yue and H. K. C., "Unified Near-Field and Far-Field Localization for AOA and Hybrid AOA-TDOA Positionings," *IEEE Transactions on Wireless Communications,* vol. 17, no. 2, pp. 1242-1254, 2018.

[67] T. Tuncer and B. Friedlander, Classical and Modern Direction-of-Arrival Estimation, 1st ed., Amsterdam, The Netherlands: Elsevier, 2009.

[68] CRFS, "Hybrid AOA/TDOA Geolocation," 9 2018. [Online]. Available: https://www.crfs.com/applicationstory/hybrid-aoa-tdoa/. [Accessed 27 12 2019].

[69] J. Ramiro Martínez-de Dios, A. Ollero, F. Fernández and C. Regoli, "On-line RSSI-range model learning for target localisation and tracking," *Journal of Sensor and Actuator Networks,* vol. 6, no. 3, p. 15, 2017.

[70] G. Vijay K., WIRELESS COMMUNICATIONS AND NETWORKING, 1, Ed., San Francisco: Morgan Kaufmann Publishers is an imprint of Elsevier, 2007, p. 48.

[71] F. Lemic, V. Handziski, M. Aernouts, T. Janssen, R. Berkvens, A. Wolisz and J. Famaey, "Regression-based estimation of individual errors in fingerprinting localisation," *IEEE Access,* vol. 7, pp. 33652-33664, 2019.

[72] S. e. a. Latre, "City of things: An integrated and multi-technology testbed for IoT smart city experiments," in *Smart Cities Conference (ISC2)*, 2016 IEEE International.

[73] H. Zhe, L. You, P. Ling and O. Kyle, “Enhanced Gaussian Process-Based Localization Using a Low Power Wide Area Network,” *164 IEEE COMMUNICATIONS,* vol. 23, no. 1, pp. 164-167, 2019.

[74] E. Quilang, T. MAKI and M. Du, “Variationi Characteristics of Meteorological Factors during Dust Storm at Dhunhuang,China,” *Faculty of Agriculture Kyushu University,* vol. 50, pp. 189-199., 2005.

[75] H. Mujlid and I. Kostanic, “Propagation Path Loss Measurements for Wireless Sensor Networks in Sand and Dust Storms,” *Frontiers in Sensors (FS),* vol. 4, pp. 33-40, 2016.

[76] G. Horvat, D. Šoštarić and D. Žagar, “Power consumption and RF propagation analysis on ZigBee XBee modules for ATPC,” Prague, 2012.

[77] T. Rappaport, Wireless communications: Principles and practice., Upper Saddle River, NJ: Prentice Hall,, 1996.

[78] S. Fadil and N. Abuhamoud, “Prediction of Microwave Signal Attenuation due to Dust and Sand Storms at (4-18 GHz)Case of study (south of Libya),” *Journal of Pure & Applied Science,* vol. 17, no. 4, pp. 1-6, 2019.

[79] M. R. ISLAM, Z. ELABDIN, O. ELSHAIKH and O. KHALIFA, “Prediction of signal attenuation due to duststorms using MIE scattering,” *IIUM Engineering Journal,* vol. 11, no. 1, pp. 71-87, 2010.

[80] U. Khan, P. Lazaridis, H. Mohamed, R. Albarracín, Z. Zaharis, R. Atkinson, C. Tachtatzis and I. Glover, "An Efficient Algorithm for Partial Discharge Localization in High-Voltage Systems Using Received Signal Strength," *Sensors,* vol. 18, no. 4000, pp. 1-19, 2018.

[81] T. ASHRAF, K. GEORGES, S. Y. V. SHAHROKH and G. FRANCOIS, "A Look at the Recent Wireless Positioning Techniques With a Focus on Algorithms for Moving Receivers," *IEEE Access,* vol. 4, pp. 6652-6680, 2016.

[82] K.-H. Lam, C.-C. Cheung and W.-C. Lee, "LoRa-based localization systems for noisy outdoor environment," Rome,Italy, 2017.

[83] thethingsnetwork, "Building a global open LoRaWAN™ network," 2016. [Online]. Available: https://www.thethingsnetwork.org/. [Accessed 3 2017].

[84] arjanvanb, "Limitations: data rate, packet size, 30 seconds uplink and 10 messages downlink per day Fair Access Policy," 2016. [Online]. Available: https://www.thethingsnetwork.org/forum/t/limitations-data-rate-packet-size-30-seconds-uplink-and-10-messages-downlink-per-day-fair-access-policy/1300. [Accessed 31 July 2019].

[85] S. AN1200.13, "LoRa Modem Design Guide," 2013.

[86] M. Kooijman, "Spreadsheet for LoRa airtime calculation," February 2016. [Online]. Available:

https://docs.google.com/spreadsheets/d/1voGAtQAjC1qBmaVuP1ApNKs1ekgUjavHuVQIXyYSvNc/edit#gid=0. [Accessed February 2020].

[87] W. Kim, W. Cho, J. Choi, J. Kim, C. Park and J. Choo, "A Comparison of the Effects of Data Imputation Methods on Model Performance," Korea (South), 2019.

[88] J. L. Schafer and J. W. Graham, "Missing data: our view of the state of the art," *Psychological Methods,* vol. 7, no. 2, p. 147–177, 2002.

[89] A. Bensky, Wireless Positioning Technologies and Applications, 2nd ed.;, 2nd ed., Norwood, MA, USA,: Artech House, 2008.

[90] C. Ning, R. Li and K. Li, "Outdoor Location Estimation Using Received Signal Strength-Based Fingerprinting," *Springerlink,* vol. 89, no. Wireless Pers Commun, p. 365–384, 2016.

[91] P. Bahl and V. Padmanabhan, "Radar: An in-building rf-based user location and tracking system.," *Proceedings of IEEE,* no. IEEE 19th Annu. Joint Conf. of the IEEE Comput. and Commun. Soc, p. 775–784, 2000.

[92] S. Xia, Y. Liu, G. Yuan, M. Zhu and Z. Wang, "Indoor Fingerprint Positioning Based on Wi-Fi: An Overview," *International Journal of Geo-Information,* vol. 6, no. 135, pp. 1-25, 2017.

[93] T. Janssen, M. Aernouts, R. Berkvens and M. Weyn, "Outdoor fingerprinting localisation using Sigfox," *International Conference on Indoor Positioning and Indoor Navigation (IPIN,* no. IEEE, pp. 1-6, 2018.

[94] L. Zhang, X. Liu, J. Song, C. Gurrin and Z. Zhu, "A Comprehensive Study of Bluetooth Fingerprinting-based Algorithms For Localization," 2013.

[95] S. Zhang, J. Guo, N. Luo, L. Wang, W. Wang and K. Wen, "Improving Wi-Fi fingerprint positioning with a pose recognition-assisted SVM algorithm," *Remote Sensing,* vol. 11, no. 6, pp. 1-23, 2019.

[96] A. Abdallah, S. Saab and Z. Kassas, "A machine learning approach for localisation in cellular environments," *Location and Navigation Symposium (PLANS),* no. IEEE/ION Position, pp. 1223-1227, 2018.

[97] B. Jingxue, W. Yunjia, L. Xin, Q. Hongxia, C. Hongji and X. Shenglei, "An Adaptive Weighted KNN Positioning Method Based on Omnidirectional Fingerprint Database and Twice Affinity Propagation Clustering," *sensors,* vol. 18, no. 2502, pp. 1-17, 2018.

[98] Q. Zhang and S. Sun, "A Centroid k -Nearest Neighbor Method," 2010.

[99] scikit-learn, "Cross-validation: evaluating estimator performance," 2009. [Online]. Available: https://scikit-learn.org/stable/modules/cross_validation.html. [Accessed 4 2019].

[100] S. Yiu, M. Dashti, H. Claussen and F. Perez-Cruz, "Wireless RSSI fingerprinting localization," *Signal Processing,* vol. 131, pp. 235-244, 2017.

[101] S. Ke, M. Zhenjie, Z. Rentong, H. Wenbiao and C. Hongsheng, "Support vector regression based indoor location in IEEE 802.11 environments," *Mobile Information Systems,* vol. 2015, pp. 1-4, 2015.

[102] S. M. Clarke, J. H. Griebsch and T. W. Simpson, "Analysis of Support Vector Regression for Approximation of Complex Engineering Analyses," *Journal of Mechanical Design,* vol. 127, no. 6, p. 1077, 2005.

[103] Z. M. Livinsa and S. Jayashri, "Localization with beacon based support vector machine in Wireless Sensor Networks," in *2015 International Conference on Robotics, Automation, Control and Embedded Systems (RACE)*, Chennai, India, 18-20 Feb. 2015.

[104] K. Rahul and A. Mariette, Efficient Learning Machines: Theories, Concepts, and Applications for Engineers and System Designers, 1st ed., Apress, 2015, pp. 39-80.

[105] T. Kleynhans, M. Montanaro, A. Gerace and C. Kanan, "Predicting Top-of-Atmosphere Thermal Radiance Using MERRA-2 Atmospheric Data with Deep Learning," *Remote Sensing,* vol. 9, no. 1133, 2017.

[106] D. Ashourloo, H. Aghighi, A. Matkan, R. Mobasheri and M. Rad, "An Investigation Into Machine Learning Regression Techniques for the Leaf Rust Disease Detection Using

Hyperspectral Measurement," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing,* vol. 9, no. 9, pp. 4344-4351, 2016.

[107] A. Omidi, S. Barakati and S. Tavakoli, "Application of nusupport vector regression in short-term load forecasting," in *2015 20th Conference on Electrical Power Distribution Networks Conference (EPDC)*, Zahedan, 2015.

[108] J. Ma, J. Theiler and S. Perkins, "Accurate on-line support vector regression," *Neural computation,* vol. 15, no. 11, pp. 2683-2703, 2003.

[109] V. Vapnik, The nature of statistical learning theory., New York: Springer, 1995.

[110] K. Shi, Z. Ma, R. Zhang, W. Hu and H. Chen, "Support vector regression based indoor location in IEEE 802.11 environments," *Mobile Information Systems,* 2015.

[111] K. P. Bennett and C. Campbell, "Support vector machines: Hype or hallelujah?," *Acm Sigkdd Explorations Newsletter,* vol. 2, no. 2, pp. 1-13, 2000.

[112] C. Kai, L. Zhenzhou, W. Yuhao, S. Yan and Z. Yicheng, "Mixed kernel function support vector regression for global sensitivity analysis," *Elsevier Ltd,* vol. 96, no. 2017, p. 201–214, 2017.

[113] C. Rasmussen and C. Williams, Gaussian Processes for Machine Learning, MIT Press, 2006, p. 13.

[114] D. Basak, S. Pal and D. Patranabis, " Support vector regression," *Neural Information Processing-Letters and Reviews,* vol. 11, no. 10, pp. 203-224, 2007.

[115] Y. Tan and J. Wang, "A Support Vector Machine with a Hybrid Kernel and Minimal Vapnik-Chervonenkis Dimension," *IEEE Transactions on Knowledge and Data Engineering,* vol. 16, no. 4, pp. 385-395, 2004.

[116] L. Rokach and O. Maimon, Data mining with decision trees: theory and applications, vol. 69, Singapore: World Scientific Publishing Co. Pte. Ltd, 2008.

[117] L. Rokach and O. Maimon, "Top-down induction of decision trees classifiers-a survey," *IEEE Transactions on Systems, Man, and Cybernetics, Part C,* vol. 35, no. 4, pp. 476-487, 2005.

[118] O. Badawy and M. Hasan, "Decision tree approach to estimate user location in WLAN based on location fingerprinting," Cairo, Egypt, 2007.

[119] A. Viel, A. Brunello, M. A. and F. Pittino, "An original approach to positioning with cellular fingerprints based on decision tree ensembles," *Journal of Location Based Services,* vol. 13, no. 1, pp. 25-52, 2018.

[120] Z. Merhi, M. Elgamel and M. Bayoumi, "A lightweight collaborative fault tolerant target localization system for wireless sensor networks," *IEEE Transactions on Mobile Computing,* vol. 8, no. 12, p. 1690–1704, 2009.

[121] A. Swetapadma and A. Yadav, "A Novel Decision Tree Regression-Based Fault Distance Estimation Scheme for Transmission Lines," *IEEE Transactions on Power Delivery,* vol. 32, no. 1, pp. 234-245, 2017.

[122] H. Ahmadi and R. Bouallegue, "RSSI-based localization in wireless sensor networks using Regression Tree," in *IWCMC 2015 - 11th International Wireless Communications and Mobile Computing Conference*, 2015.

[123] V. Andrea, B. Andrea, M. Angelo and P. Federico, "An Original Approach to Positioning with Cellular Fingerprints Based on Decision Tree Ensembles," *Springer International Publishing,* vol. AG, pp. 49-70, 2018.

[124] T. Dietterich, "Ensemble Methods in Machine Learning," *Springer,* vol. 1857, pp. 1-15, 2000.

[125] L. Breiman, "Random Forests," *Machine Learning,* vol. 45, no. 1, pp. 5-32 , 2001.

[126] T. K. Ho, "Random Decision Forests," *Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC,* pp. 278-282, 14-16 August 1995.

[127] Y. Chen, M. Sakamura, J. Nakazawa, T. Yonezawa, A. Tsuge and Y. Hamada, "OmimamoriNet: an outdoor positioning system based on Wi-SUN FAN network.," *Eleventh International Conference on Mobile Computing and Ubiquitous Network (ICMU) ,* no. IEEE, pp. 1-6, 2018.

[128] Z. Yanru and H. Ali, "A gradient boosting method to improve travel time prediction," *Elsevier-Transportation Research Part C* , no. 58 , p. 308–324, 2015.

[129] F. Lemic, V. Handziski, M. Aernouts, T. Janssen, R. Berkvens, A. Wolisz and J. Famaey, "Regression-based estimation of individual errors in fingerprinting localisation," *IEEE Access,* vol. 7, pp. 33652-33664, 2019.

[130] M. Praveena and V. Jaiganesh, "A Literature Review on Supervised Machine Learning Algorithms and Boosting Process," *International Journal of Computer Applications,* vol. 169, pp. 32-35, 2017.

[131] R. Barzegar, A. Moghaddam, J. Adamowski and B. Ozga-Zielinski, "Multi-step water quality forecasting using a boosting ensemble multi-wavelet extreme learning machine model," *Stoch Environ Res Risk Assess,* p. 799–813, 2018.

[132] P. Verma and R. Kumar, "A Literature Survey on Classification Algorithms of Machine Learning," *International Journal of Computer Applications,* vol. 179, pp. 47-50, 2018.

[133] L. Breiman, "Arcing The Edge," *Technical Report 486, Statistics Department, University of California at Berkeley,* 1997.

[134] R. Caruana and A. Niculescu-Mizil, " An Empirical Comparison of Supervised Learning Algorithms," *Proceedings of the 23rd International Conference on Machine Learning,* pp. 25-29, 2006.

[135] T. Hastie, R. Tibshirani and J. Friedman, "10.10 Numerical Optimization via Gradient Boosting," in *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed., New York, Springer, 2009, pp. 358-361.

# APPENDIX 1 EPSILON-SVR

Appendix 1 is a collection of the results obtained for parameter optimisation in the kernel functions used in the epsilon-SVR algorithm for developing node localisation models in Chapter 5 of this thesis. The matern Kernel which combined with other kernels, has a "nu_Matern" parameter that is needed to be optimised. Random search method was used to select the optimal "nu_Matern" parameters.

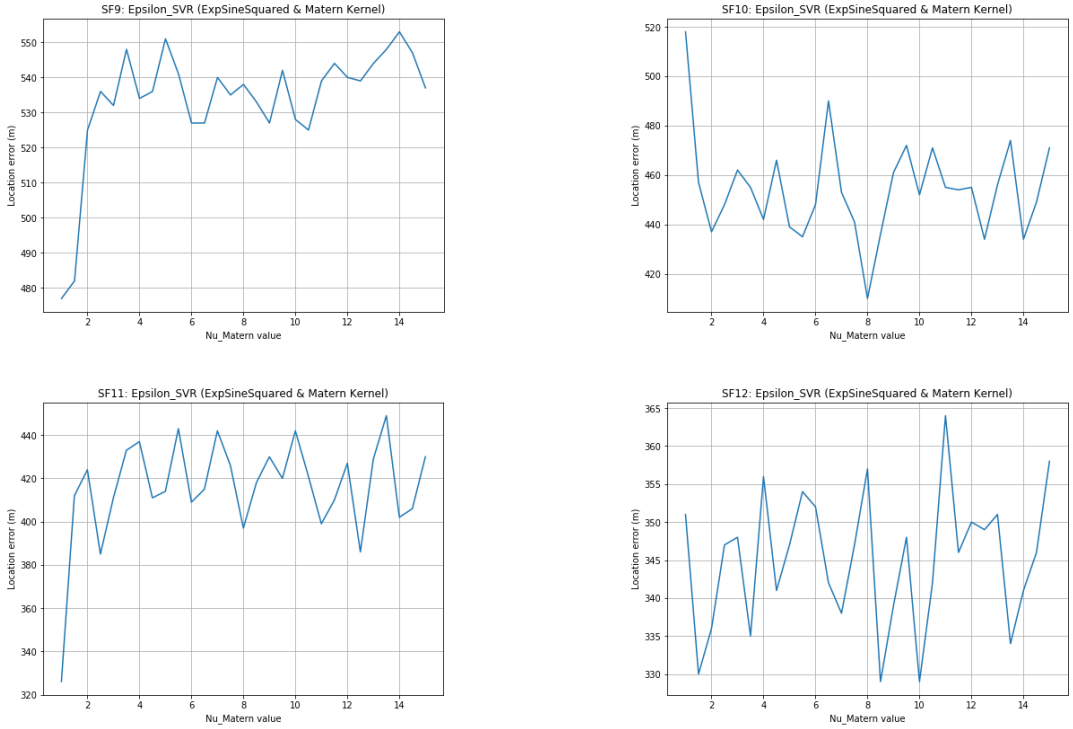



**Figure A1.1: ExpSineSquared and Matern kernels**

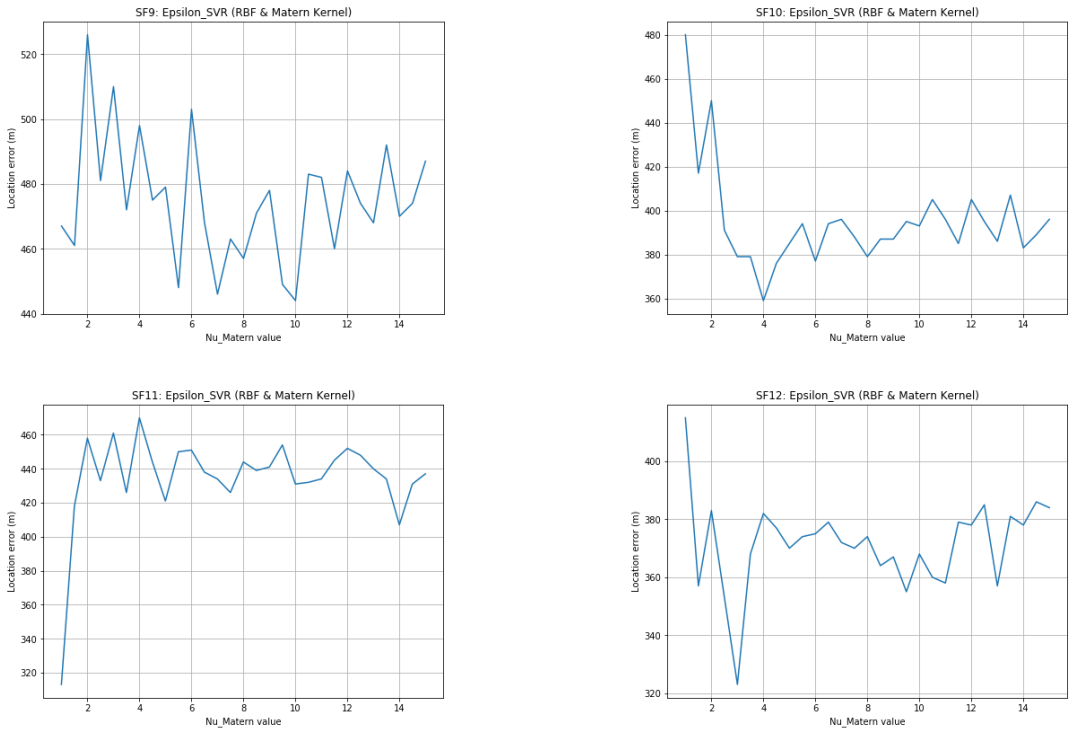



**Figure A1.2: RBF and Matern Kernel**

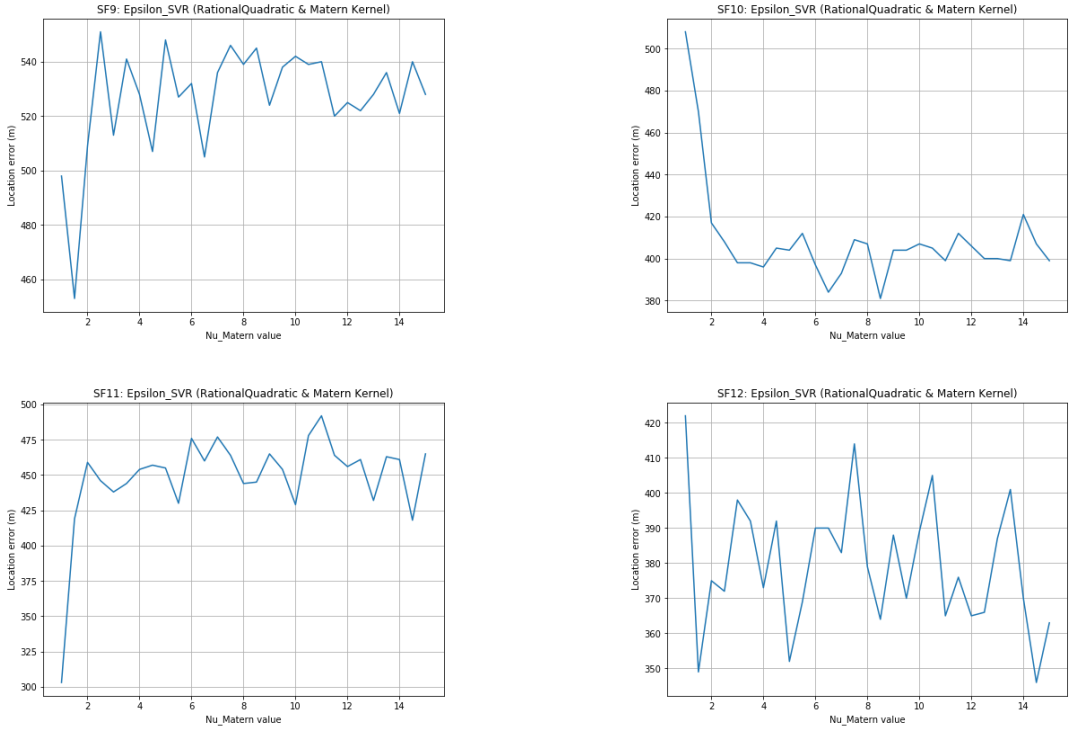



**Figure A1.3: RationalQuadratic + Matern kernel**

# APPENDIX 2 NU_SVR

Appendix 2 presents plots obtained in the process of parameter optimisation in both the nu-SVR algorithm and the kernel functions used in developing node localisation models in Chapter 5, of this thesis. Random search method was used to tune the parameters of matern Kernel and nu-SVR algorithm. The minimum mean values in the plots represent the optimal value of each parameter.

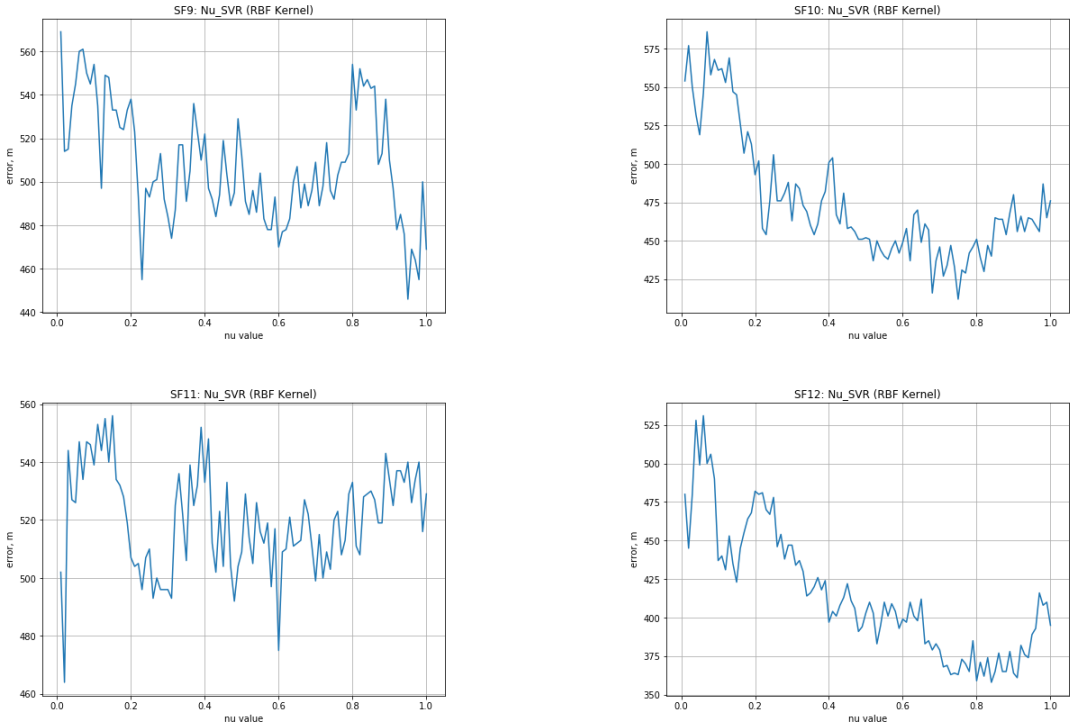



**Figure A2.1: RBF Kernel**

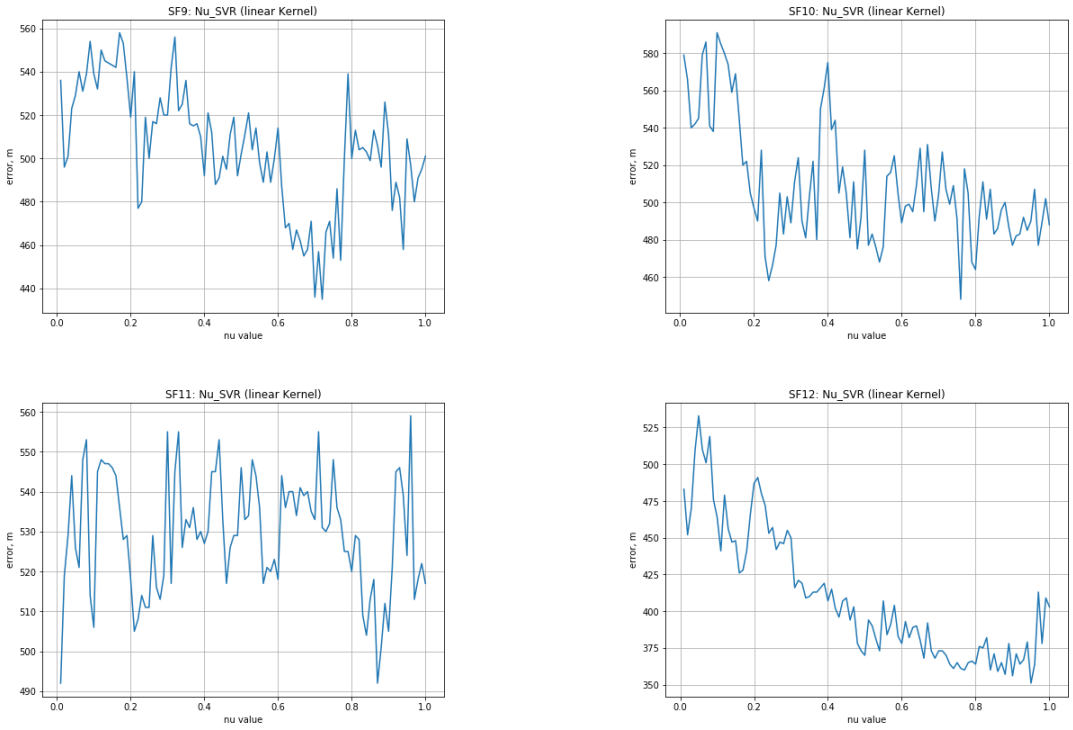



**Figure A2.2: Linear Kernel**

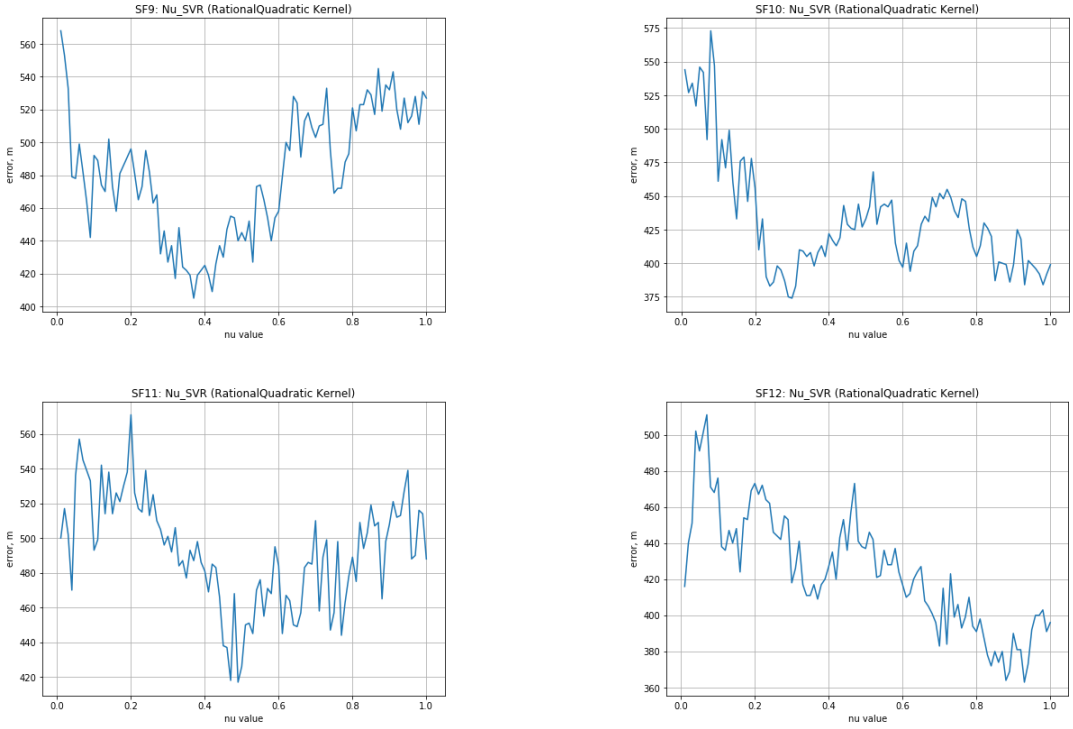



**Figure A2.3: Rational Quadratic Kernel**

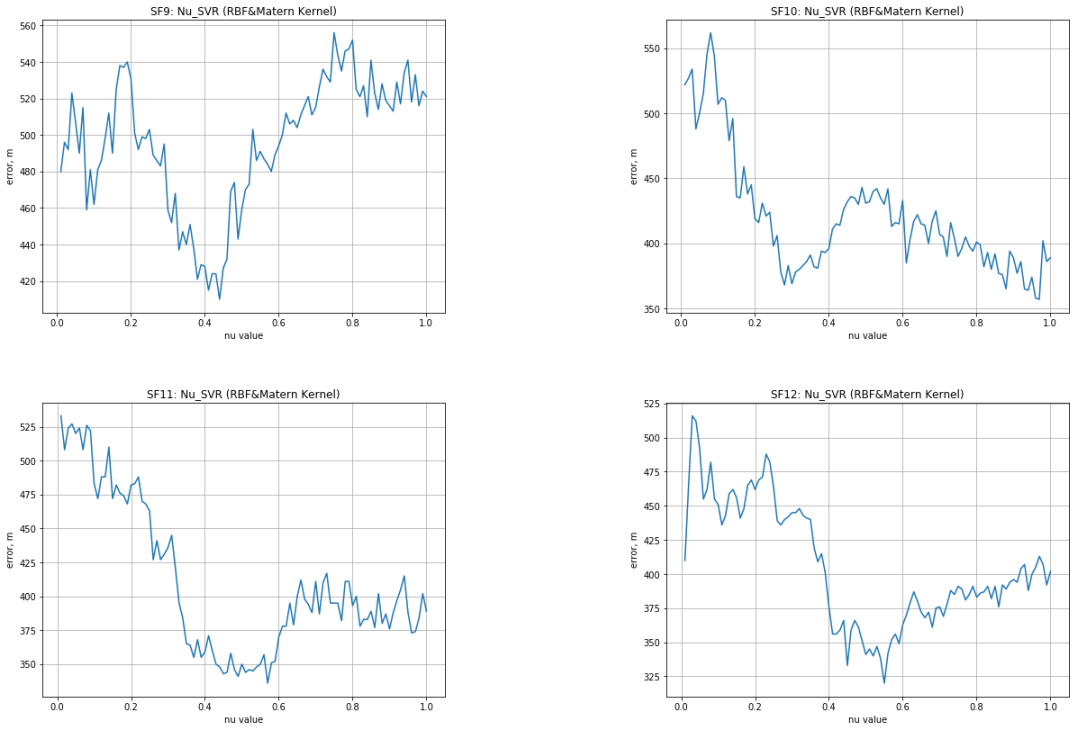



**Figure A2.4: RBF+Matern**

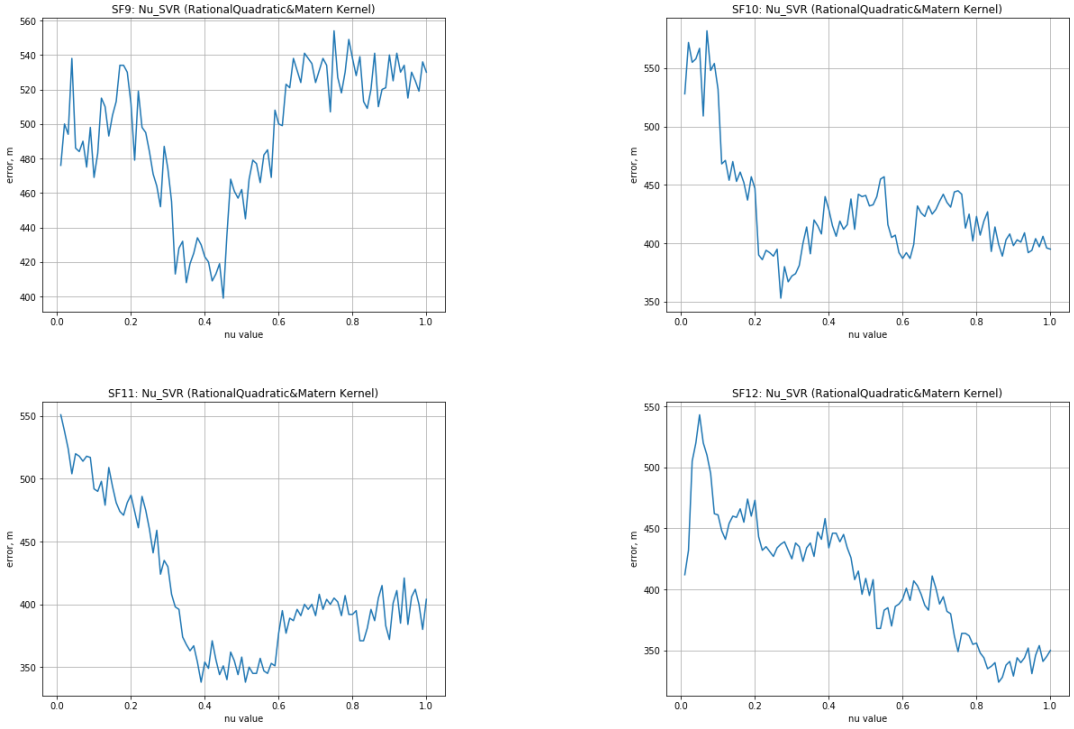



**Figure A2.5: Rational Quadratic + Matern**

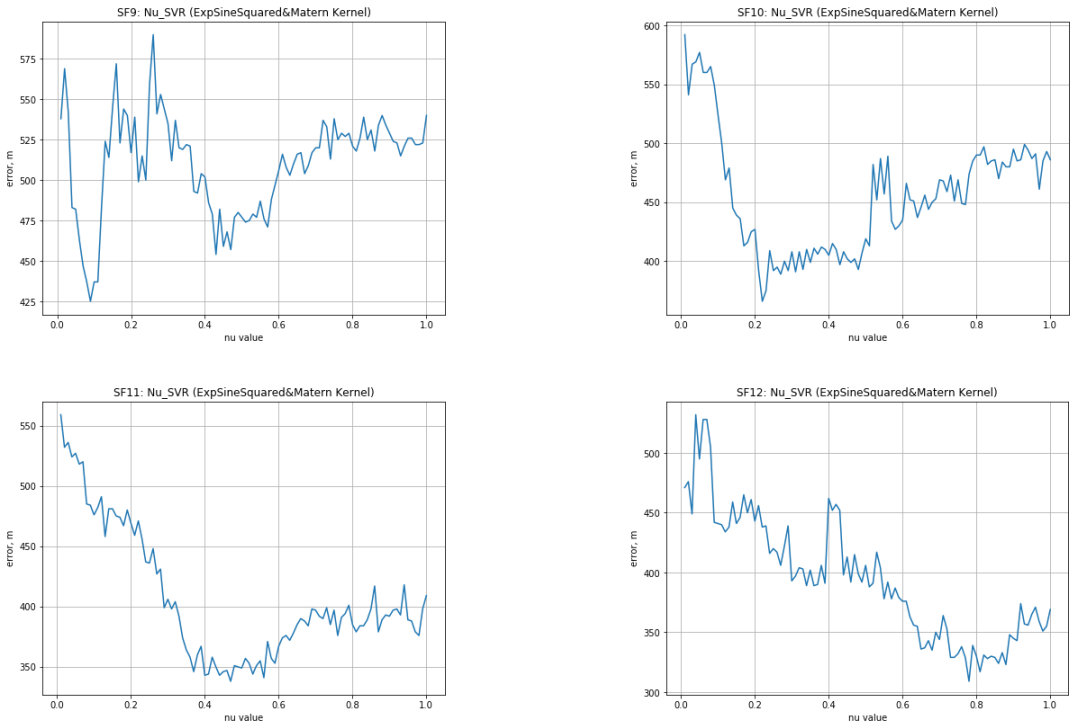



**Figure A2.6: ExpSineSquared + Matern**

# APPENDIX 3 GAUSSIAN PROCESS REGRESSION

Appendix 3 is a collection of the plots that results from the process of parameter optimisation in the kernel functions used for developing Gaussian Process Regression based node localisation models in Chapter 5 of this thesis. Random search method was employed to tune the parameters and the optimal parameter was given has the minimum mean value in the obtained plots shown below.

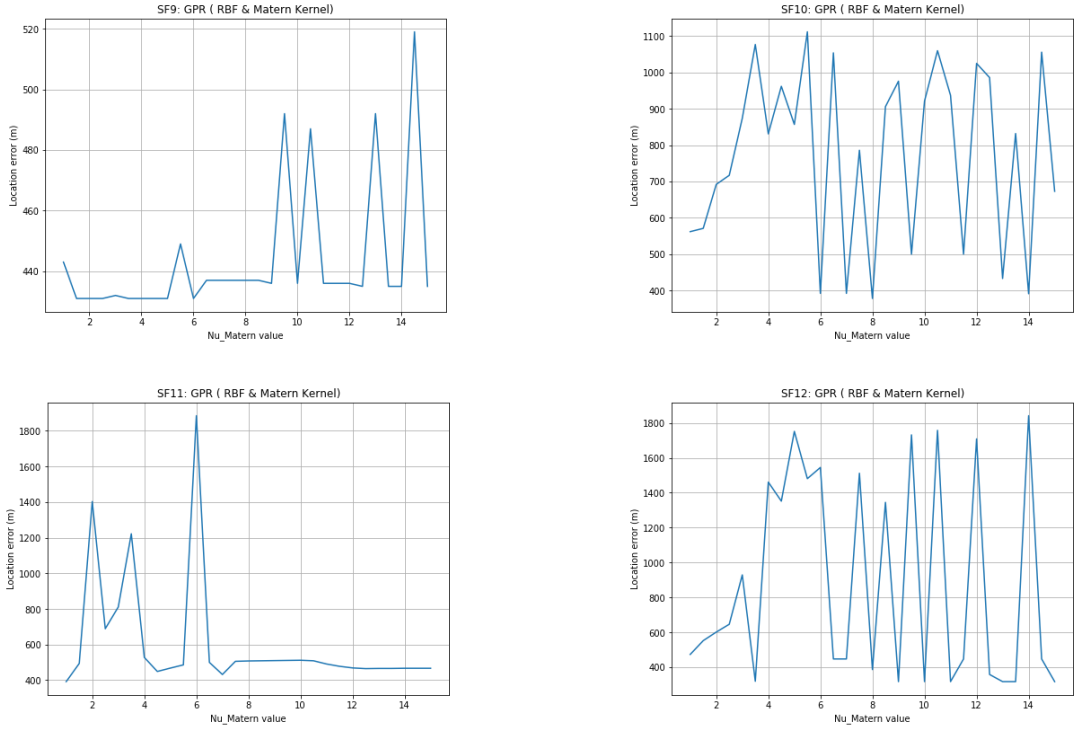



**Figure A3.1: RBF & Matern kernels**

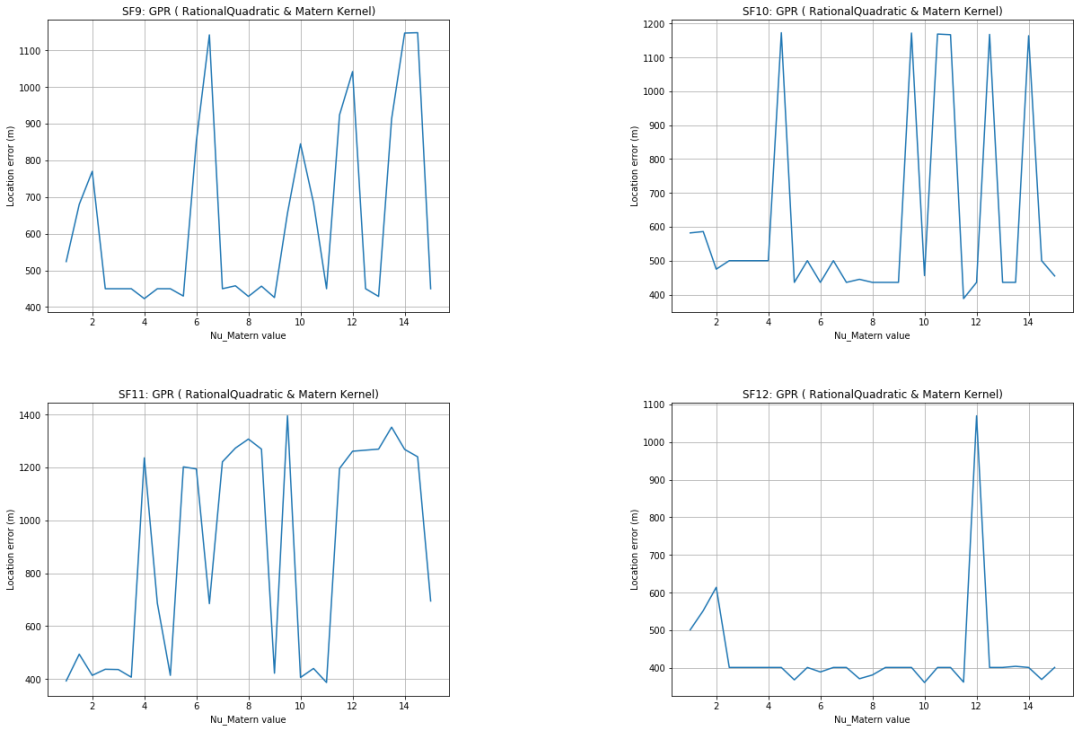



**Figure A3.2: Rational Quadratic & Matern kernels**

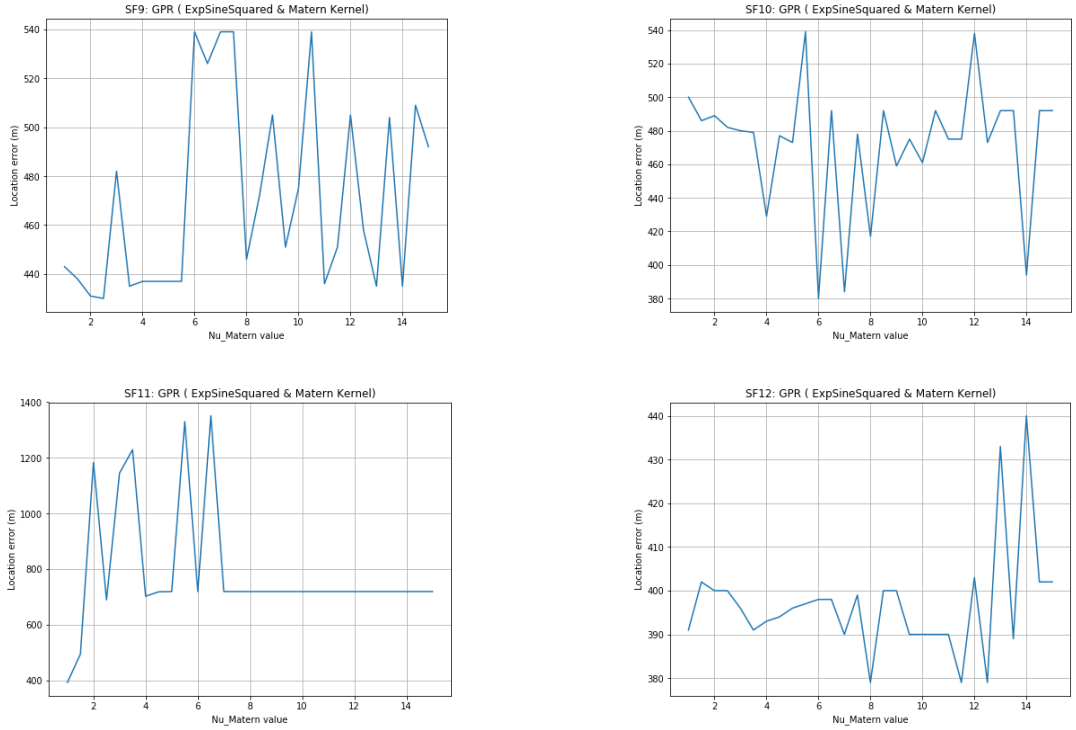
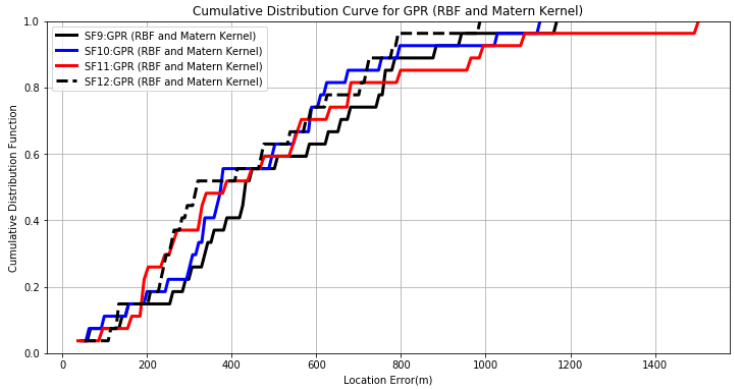



**Figure A3.3: ExpSinSquared & Matren kernels**

# APPENDIX 4 CDF AND BOXPLOTS FOR KERNEL-BASED MODELS

Appendix 4 is a collection of the obtained results for the analysis of the developed kernel-based node localisation models in Chapter 5 of this thesis. Both the Cumulative Distribution Frequency (CDF) and the boxplots represent the performance the individual models using different kernel functions. This gives the overall performance of each model in terms of localisation error.



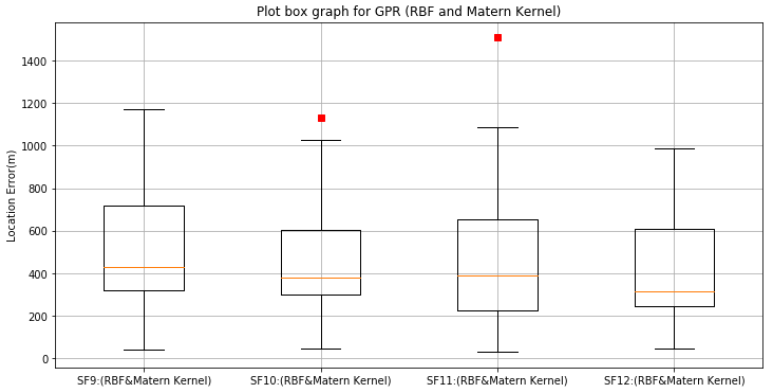
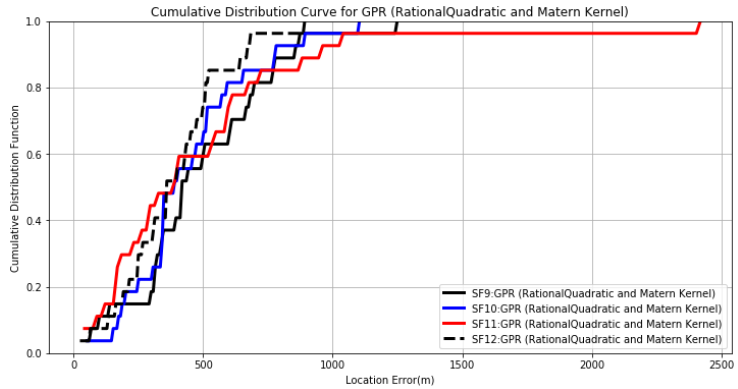




**Figure A4.1: CDF and box plot for GPR using RBF + Matern kernel.**



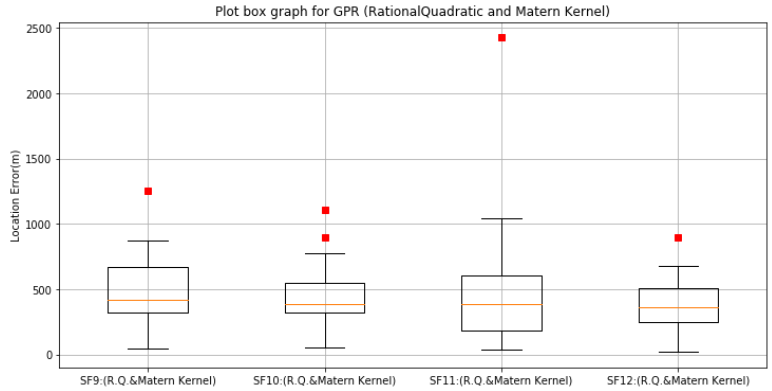
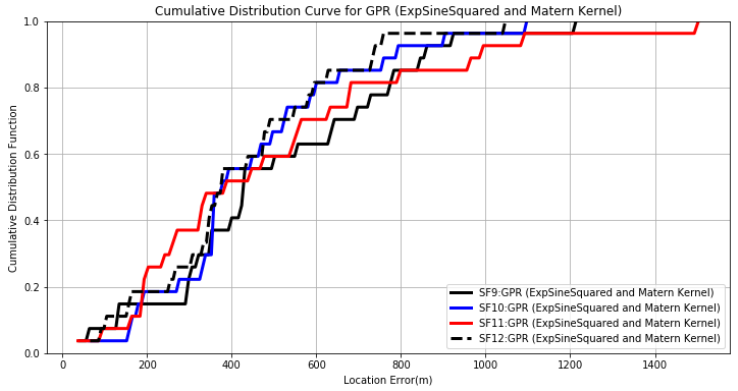




**Figure A4.2: CDF and box plot for GPR using Rational Quadratic + Matern Kernel.**



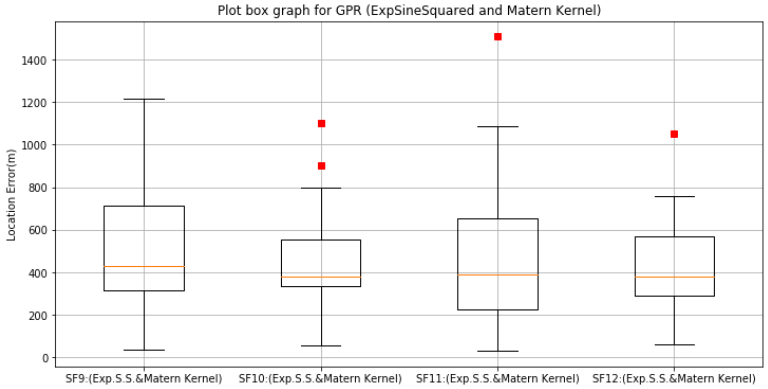
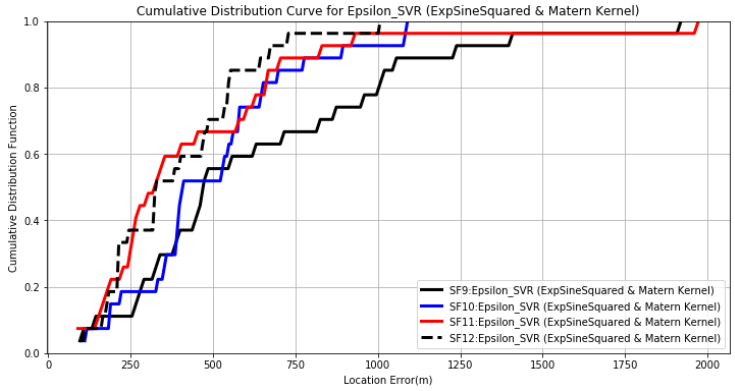




**Figure A4.3: CDF and box plot for GPR using ExpSineSquared + Matern Kernel Function.**



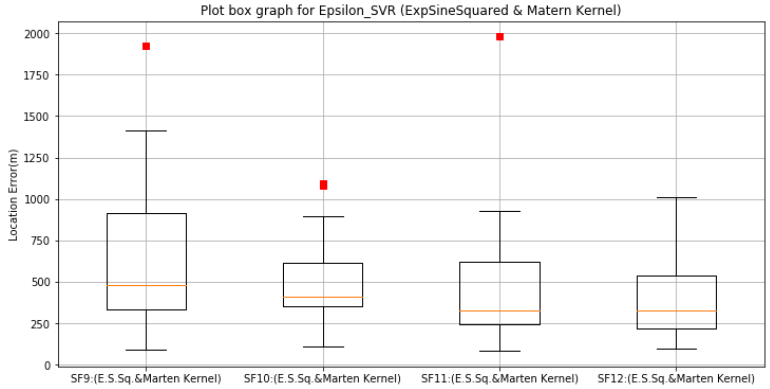




**Figure A4.4: CDF and box plot for epsilon-SVR using ExpSineSquared + Matern Kernel Function.**

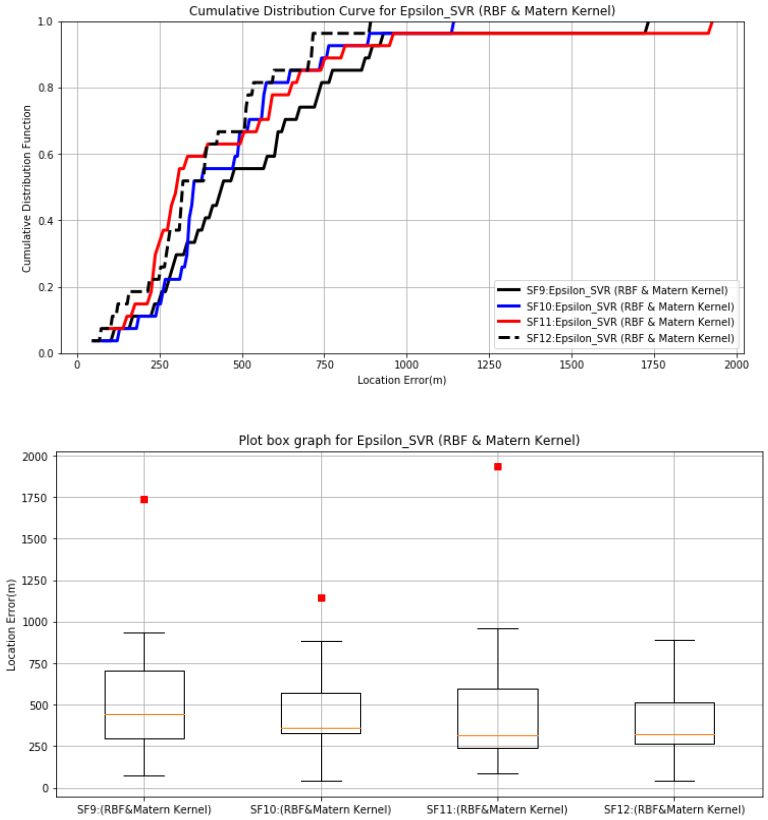



**Figure A4.5: CDF and box plot for epsilon-SVR using RBF+Matern Kernel.**

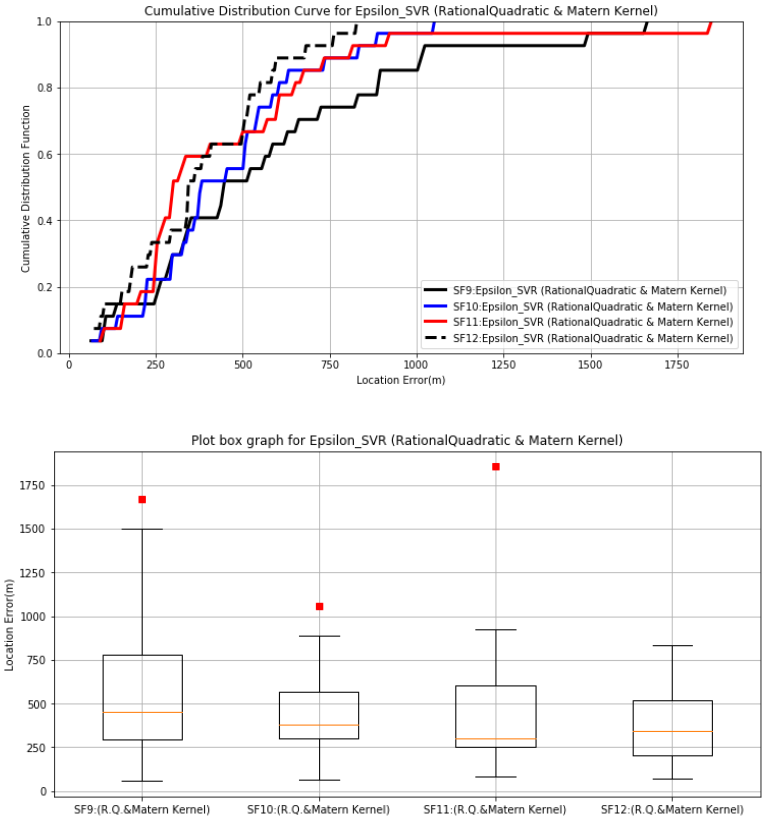



**Figure A4.6: CDF and box plot for epsilon-SVR using Rational Quadratic + Matern Kernel.**

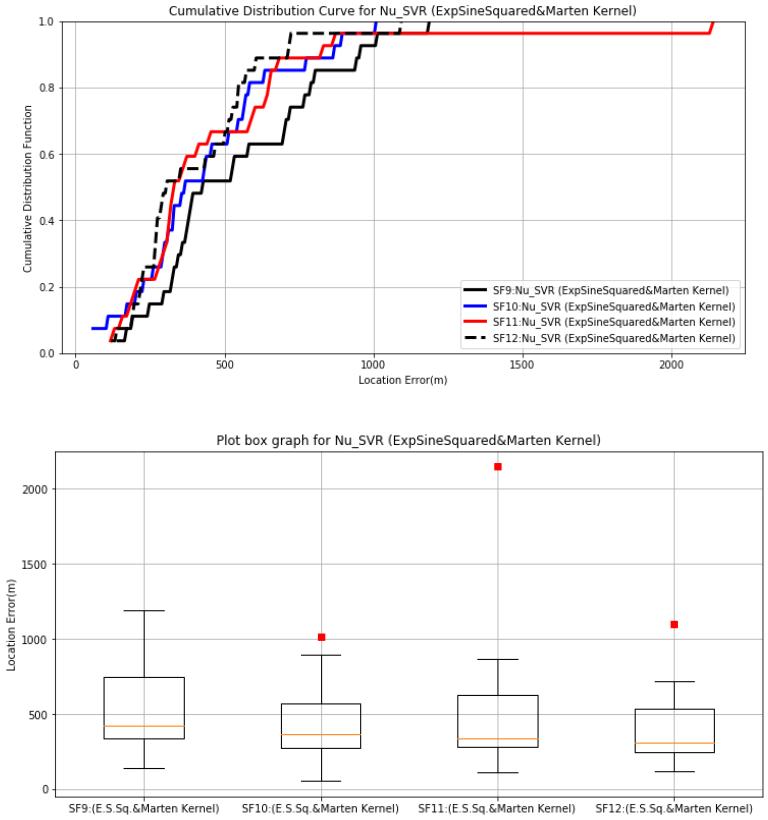



**Figure A4.7: CDF and box plot for nu-SVR using ExpSineSquared + Matern Kernel Function.**

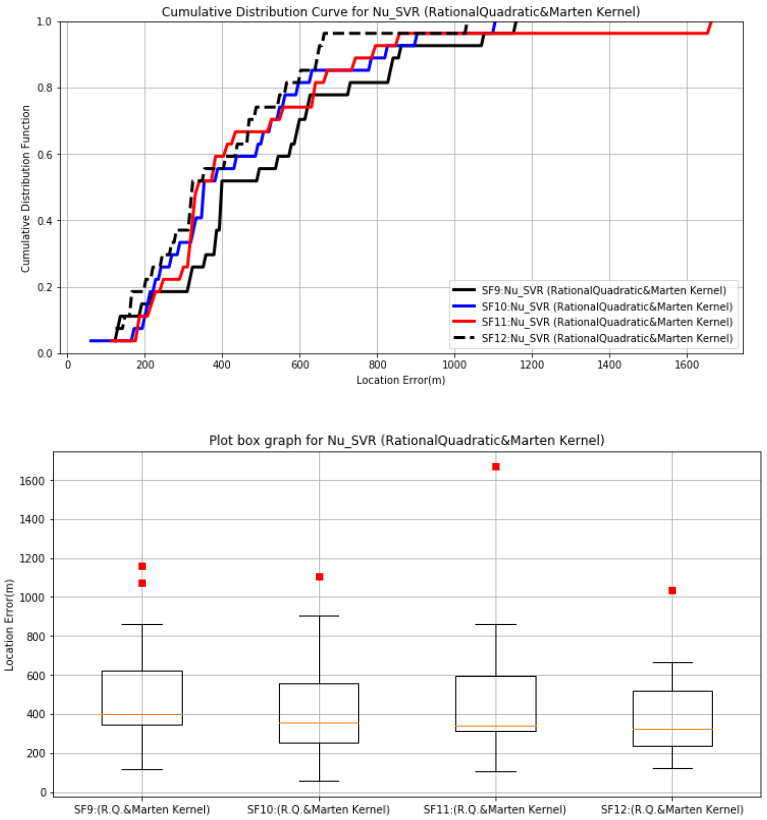



**Figure A4.8: CDF and box plot for nu-SVR using Rational Quadratic + Matern Kernel.**

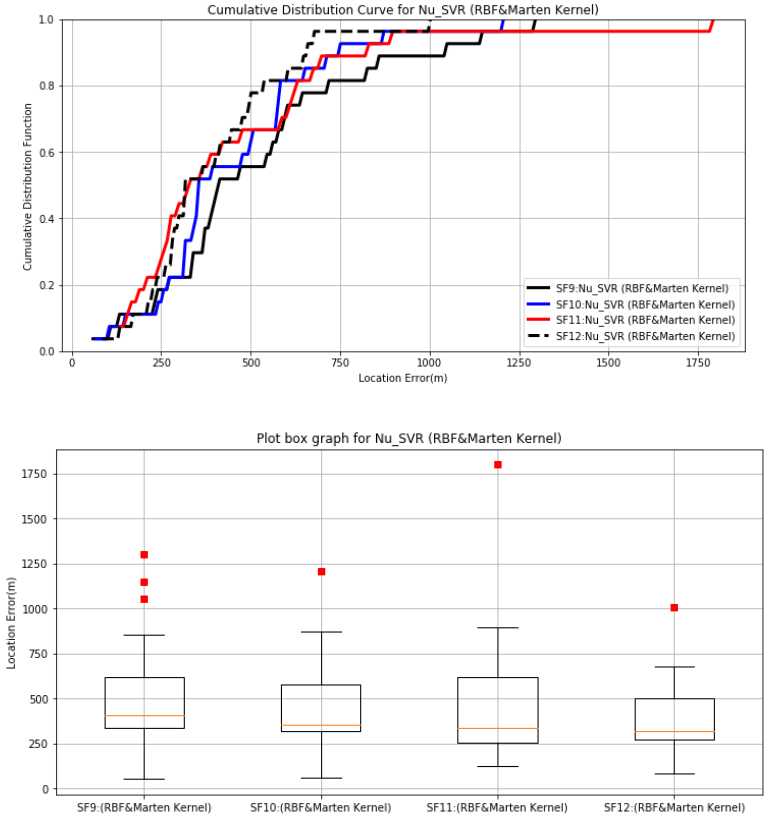



**Figure A4.9: CDF and box plot for nu-SVR using RBF + Matern Kernel.**