STATISTICAL INFERENCE FOR PERIODIC AND PARTIALLY OBSERVABLE POISSON PROCESSES

by

FERDIAN JOVAN

A thesis submitted to
The University of Birmingham
for the degree of
DOCTOR OF PHILOSOPHY

School of Computer Science
College of Engineering and Physical Sciences
The University of Birmingham
April 2018

Abstract

This thesis develops practical Bayesian estimators and exploration methods for count data collected by autonomous robots with unreliable sensors for long periods of time. It addresses the problems of drawing inferences from temporally incomplete and unreliable count data.

This thesis contributes statistical models with spectral analysis which are able to capture the periodic structure of count data on extended temporal scales from temporally sparse observations. It is shown how to use these patterns to i) predict the human activity level at particular times and places and ii) categorize locations based on their periodic patterns.

The second main contribution is a set of inference methods for a Poisson process which takes into account the unreliability of the detection algorithms used to count events. Two tractable approximations to the posterior of such Poisson processes are presented to cope with the absence of a conjugate density. Variations of these processes are presented, in which (i) sensors are uncorrelated, (ii) sensors are correlated, (iii) the unreliability of the observation model, when built from data, is accounted for. A simulation study shows that these partially observable Poisson process (POPP) filters correct the over- and under-counts produced by sensors.

The third main contribution is a set of exploration methods which brings together the spectral models and the POPP filters to drive exploration by a mobile robot for a series of nine-week deployments. This leads to (i) a labelled data set and (ii) solving an exploration-exploitation trade-off: the robot must explore to find out where activities congregate, so as to then exploit that by observing as many activities.

ACKNOWLEDGEMENTS

First, I would like to express my innermost gratitude to my amazing supervisors, Jeremy Wyatt and Nick Hawes. Without their encouragement, motivation and wisdom, I would not be like the person I should be. Not only did they support my research and my self-actualisation, but also gave me enough freedom to enjoy my life during my PhD. They taught me patiently and helped me get out of a mud when I was stuck. More importantly, they tiredlessly motivate and guide me to find self-assurance with what I was doing.

A very special gratitude goes to the STRAND Project which funded my research. Without this funding and the help of Jeremy Wyatt and Nick Hawes, pursuing my PhD would be impossible. Also, many thanks must go to Peter Tino and Ela Claridge for their insightful comments as members of my thesis group.

I thank all STRANDS team and especially to those who worked close to me: Dr. Lars Kunze, Dr. Chris Burbridge, Dr. Bruno Lacerda, Dr. Lenka Mudrova, Dr. Paul Duckworth, Dr. Mohannad Omari. Thank you for your support in numerous obstacles I have been facing through my research. My research which is part of this project became a big success because all of you. I would like to be honest that my experience in completing this project is wholly fun and enjoyable. My special thanks algo goes to Milan Tomy for her help during the last 6 months of my final experiment. Also, for our precious robots, Bob and Betti for helping me running my experiments.

My sincere thanks also to all of my friends in computer science: Syahril Shafie, Adhi Wicaksana, I Made and his family, Sean Bastable, Marco Bercerra, Fatma Faruq, Xiaodong, for all hang outs, advices, cries and laughter. Thank you to my flatmates in 15 North Road: Rainer Schutz, Adrian Firdaus, and Luca Danieli, for making my life experience in Birmingham wonderful. My badminton friends: Surya, Alam, Keat, Andy, and my friends in Dresden: Kelvin, Eric, Adhi and his family, Christina, Ryan and Zamzami. Thank you all for being with me when I was down, thank you for all amazing laughs that we share. I can not list all the names individually, but if I have spoken to you in the last four years, then you have my gratitude for helping me through this.

Nobody has been more important to me than my parents. If it is not because of them, I would not be here pursuing my degree if they did not encourage me to be a better man. They are my spirit to be here to make them proud and happy in their old life. Most importantly, I wish to thank my wife, who provides me her encouragement during my research and my writing process. I deeply thank her for her unwavering support.

※ ※ ※ ※ ※

CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Autonomous mobile robots that work in human-centred environments are seen as a promising future application area for robotics systems. These systems are expected to leave carefully controlled industrial environments and come to assist us with menial tasks in our homes and offices. Possible applications include, for example, domestic robot assistants which help people to live independently for longer Feil-Seifer & Matarič (2011), delivery robots in hospitals, stock-keeping robots in supermarkets and warehouses, and security robots in factories.

Having robots operate in human-populated environments requires modelling human activities. This is because any tasks or plans the robots have must take into account human activities. Since human activities involve many complex interactions, they can be modelled at many levels of detail. These can range from recognising simple activities over a few seconds, a minute, or an hour; predicting what a person is going to do next; to determining whether a group of people is performing an activity together. In any scenario where a robot learns about human activities, it must first observe them. Thus, the first thing the robot needs to do is to plan to go to where people are. This problem of finding and observing people is the basic motivation for this thesis.

To be where people are, the robot must first know when and where it is likely to see people. It becomes, however, a challenging problem if one tries to predict exactly where a particular individual will be, so as to observe that individual. There are two reasons for this. First, each individual roaming in a human-populated environment is hard to re-recognise. Second, individual persons often travel long distances and visit places robots cannot follow. Hence, instead of predicting where an individual will be, it would be easier to predict where the robot is likely to see many people, without regard for exactly who it may observe. This formulation would allow the robot to observe a sufficient amount of human activity, so as to learn the specific kinds of activity models mentioned previously.

The problem of this thesis can be loosely formulated as how to predict where many people are most likely to be and to go and observe them. Specifically, it requires the robot to go to where the aggregate level of human activity is highest. In addition, this thesis chooses to tackle the problem for the case where the robot runs for an extended period of time such days or even weeks as it builds its models.

To tackle the formulated problem, a mobile robot must know not only where people are, but also when they are in those locations. It also needs a model which predicts how many people the robot will be likely to see in a particular place at a particular time since people tend to move around from one location to another.

An important restriction on using a mobile robot is that it can only be in one place at one time, so its observations are spatially restricted. As the robot moves around, it will only see particular locations infrequently. Thus its data for those locations will be temporally sparse. This adds yet another requirement to the problem formulation, where the robot must know how uncertain it is about how many people might be in one place at a particular time. This requirement is necessary since the robot has a limit to its operational life. We would, therefore, like it to optimise the time it takes to build its models. This introduces an exploration-exploitation trade-off problem Wyatt (1998), Alba & Dorronsoro (2005), Audibert et al. (2009), i.e. should the robot visit a familiar place, where it will probably see two people, or go to a new place, where it might see many more but might see no-one? Another important

restriction in mobile robotics is that robot sensors are unreliable. Any solution must take into account the unreliability of sensors. We may also require that it do so for when multiple sensors are involved.

To solve the formulated problem, any possible existing structure that governs human activities must be considered and exploited. One particular structure that aggregate human activity levels tend to have is periodicity. This periodic structure can be daily, weekly, or even hourly. For example, during term time, students frequently come to class in the morning, leave and enter lecture theatres around the hour, populate canteen areas for lunch, and leave campus before evening.

In summary, this thesis tries to solve a problem that deals with unreliable data coming from multiple sensors which represent aggregate human activity levels in various places,

- where the data, which is considered temporally incomplete, has some periodic structure, and
- where the robot must choose how to explore its environment so as to observe the maximum possible number of activities over time.

This thesis proposes a combination of signal processing and statistical models to tackle the problem. The proposed method is validated in simulation, and on real robot runs over periods of several months. Despite the application to mobile robotics, the proposed methods should be applicable to many other fields.

1.1 Sensors' Systematic Errors

In this thesis, count data represent the aggregate human activity levels. Where these counts are made using sensor data, all sensor detectors embedded in our autonomous mobile robot have some level of unreliability. The unreliability of the detectors introduces *systematic errors* in the activity datasets leading to bias in the statistical estimates.

Systematic errors are typically constant and proportional to the true value. As a result,

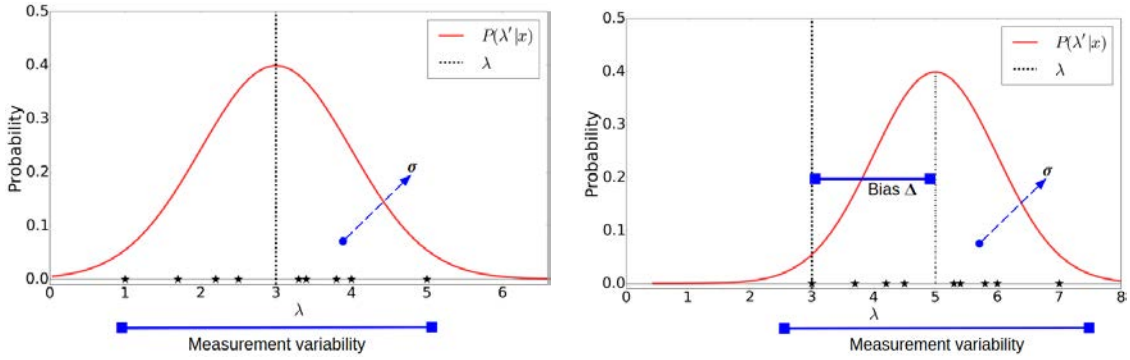

Figure 1.1: A graphical view of measurement errors. The dispersion of the distribution determines the measurement variance (left figure). The offset between the observed value and the true value is bias of measurement – systematic error (right figure)

the mean of repeating samples will never approach the true mean. Systematic errors are caused by the imperfection of the detector, or interference of the environment with the detection process, and always affect the samples in a predictable direction. Consider Figure 1.1. The variance, σ^2, describes a typical spread in the samples, and hence it represents how samples fluctuate around the mean value. Under systematic errors, samples fluctuate at random around λ' which has Δ distance from the true value λ.

In count data representing human activity levels which are made using sensor detectors, systematic errors are introduced in each bit, i.e. presence or absence of a person, that makes up the count. Each bit may experience misclassification and the count resulted from that may end up undercounting or overcounting. The count may also end up having the correct value but the bits making up the count experience mis-classification which counteracts one another. For this case, even though the count is correct and it does not show any systematic bias as a whole, the systematic bias can still be found on each bit that makes up the count. Figure 1.2 show some examples of systematic errors experienced by the perception algorithms used in this thesis for detecting humans.

Any *naive* statistical analysis will treat the error-corrupted variable to be the same as the true variable of interest and ignore the bias effect of the systematic error in the parameter estimation. An incorrect estimate might lead to incorrect inferences. An example of such an occasion in count data is the detection of differentially expressed genes, where even subtle

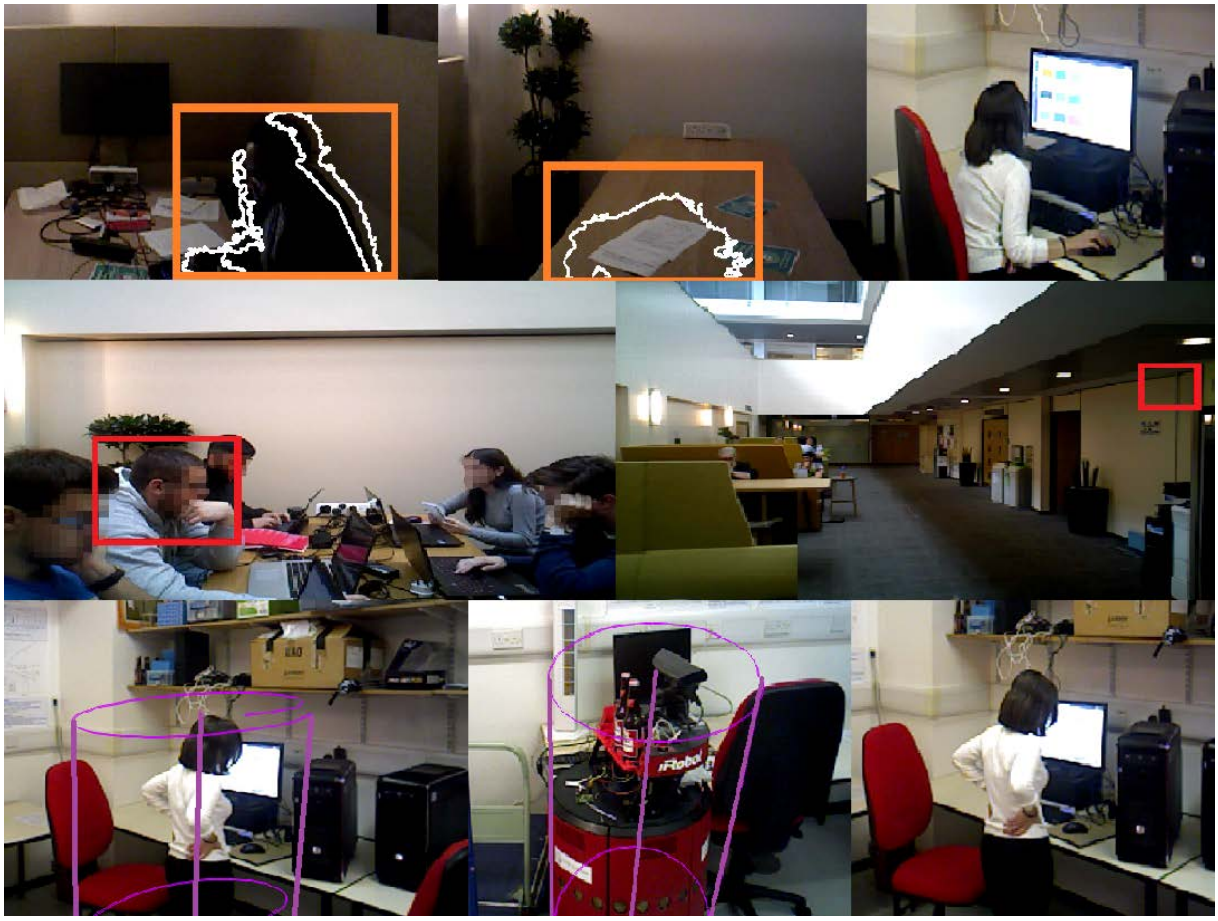

Figure 1.2: Systematic errors from different areas for each detector. Top row: change detector. Middle row: upper body detector. Bottom row: leg detector. Detections are marked with 2D or 3D bounding boxes.

changes in gene expression levels can be indicators of biologically crucial process Varuzza et al. (2008).

Despite the harmful effect that miscounted (under-or over- counted) data due to biased perception has on the estimation process, correct estimates of the parameter can still be obtained if one has an idea of the magnitude and the direction of the perceptual bias Mwalili (2006). This knowledge can be obtained from experts and be combined in a Bayesian way to correct for the bias. The same correction can also be achieved if ground-truth data are available. By statistically modelling the accuracy of a detector a correction can be made in either a frequentist or a Bayesian framework. Hence, this will be another topic of this thesis where the problem will be specifically applied to count data collected by mobile robots.

1.2 Assumptions of the Thesis

Recall that the motivation for this thesis is to create a long-term understanding of the temporal dynamics of the aggregate level of human activity, along with the ability to exploit the uncertainty and incompleteness of this to drive exploration by an autonomous mobile robot. As the motivation is quite broad, some constraints are imposed to fit the motivation into a practical robot application. Here are the following constraints and assumptions.

1. The environment is manually segmented into regions/areas. Thus there is no attempt to generalise across continuous space.

2. For each area, the level of aggregate activity is represented as count data drawn from a single non-homogeneous Poisson process.

3. The activities are observed through sensors/perception algorithms/counting algorithms embedded in an autonomous mobile robot. These sensors each observe the same underlying process.

4. Data are temporally sparse across time.

The second constraint is altered specifically for Chapter 6 to Chapter 9. Instead of assuming the count data are drawn from a single non-homogeneous Poisson process, the chapters assume that the count data are drawn from a single homogeneous Poisson process.

1.3 Contributions and Related Publications

This thesis is split into two parts. The first part focuses on developing a statistical model for count data which draws inferences from temporally sparse observations. Given a perception algorithm (or a sensor) is available and attached to a mobile robot that patrols around perimeters of a large area, it adds an assumption that any perception algorithm used to capture the observations is reliable. As a result, it reasons only about *when* and *where* the fluctuations of human activities are likely to happen. This thesis makes the following contributions.

1. It presents the development of a statistical model termed the spectral-Poisson model. This is able to capture the periodic structure of count data over an extended temporal scale from temporally sparse observations. Based on this model, the robot is able to predict the level of aggregate activity at some future time and place. The thesis demonstrates how the proposed model performs relative other temporal modelling methods. This work has been published in Jovan et al. (2016).

2. It applies the model of human activity levels developed above to unsupervised categorization of locations. The resulting categories provide a broadly similar semantic grouping to that given by humans. This work has also been published in Jovan et al. (2016).

3. It applies the proposed model to drive exploration to search for humans, specifically to maximise the aggregate levels of human activity observed. Experiments were carried out on different mobile robots in long-term autonomous deployments. The exploration model sets a series of target locations for the robot to visit at particular times of the day. Additionally, the activity-driven exploration software has been made open-source [1]

The second part of this thesis focuses on formulating what is termed a partially observable Poisson process (POPP) and developing a set of inference methods for this process. In particular, a series of approximate Bayesian filters are built to address the problem of how to efficiently correct miscounts made by either single or multiple unreliable counting devices. The second part, apart from Chapter 10, limits the case of Poisson processes to homogeneous Poisson processes. The thesis makes the following contributions.

1. It presents a set of inference methods for a Poisson process, which takes into account the unreliability of perceptual algorithms that count events. Unlike Bayesian estimation for a fully observable Poisson process (FOPP), obtaining the posterior is non-trivial, since there is no conjugate density for a POPP and the posterior has a number of elements that grow exponentially in the number of observed intervals. Two simple, tractable

[1] https://github.com/strands-project/strands_exploration

approximations, have been presented. These two approximations are combined in a switching filter which enables efficient and accurate estimation of the posterior. This work has been published in Jovan et al. (2018).

2. It presents formulations of three variations of the basic POPP filter. The POPP-Beta filter extends the POPP filter to the case in which the unreliability of the observation model is accounted for during inference. The N-POPP filter extends the POPP filter by modelling the case when sensors are uncorrelated. The POPP-Dirichlet combines the POPP-Beta filter and the N-POPP filter to have the benefits of each correction. A simulation and extended robot-gathered observations have shown that each extension provides a more accurate estimation than the POPP filter.

3. It presents the results of applying the proposed model in combination with the Spectral-Poisson to drive activity exploration on a mobile robot for a series of two-week deployments, where the model sets out a series of places to visit during work hours. The proposed model aims to accurately correct miscounts from non-homogeneous Poisson processes while the Spectral-Poisson model aims to capture the periodic structure of temporally sparse count data. The deployments also resulted in a labelled dataset of human detections in a total of six weeks of data.

One should note that the publications Jovan et al. (2016) and Jovan et al. (2018) were fully contributed by the work detailed in this thesis. Hence, the author of this thesis fully contributed to those publications.

CHAPTER 2

REVIEW OF STATISTICAL INFERENCE

The general concern of this thesis is how agents should draw inferences from observations gathered under conditions of incompleteness and with unreliable sensors. As stated in the previous chapter, this thesis is specifically concerned with the Bayesian approach to statistical inference for counting processes. Therefore, this chapter is dedicated to reviewing the general concepts and theories on which it builds. Even though all of these theories have been reviewed extensively elsewhere Carlin & Louis (2010), Cameron & Trivedi (2013), Papoulis & Pillai (2002), Fink (1997), Scott (1998), DeGroot & Schervish (2012), Kingman (2005), Tijms (2003), Karlin (2014), it will be easier for the reader to understand the rest of this thesis if the ideas and the notation employed are described and unified.

2.1 Statistical Inference

Almost all of the chapters in this thesis heavily rely on a Bayesian approach to statistical inference. Before going further into more complex definitions, let's start with the definition of statistical inference. *Statistical inference* is the process of understanding how particular data are generated. The data themselves can be thought of as being generated by a product of

some data generation processes which are unknown and hidden from an observer perspective. The hidden processes are represented by a statistical model, assuming parametric inference where the model takes a form of a probability distribution. Let $L(.)$ be the chosen model or the chosen probability distribution with parameters $\overrightarrow{\theta} = (\theta_1, \ldots, \theta_n)$ which govern the distribution, then the process of how the data are generated can be represented as:

$$X \sim L(x \mid \overrightarrow{\theta})$$

where X is a random variable representing the generated data. Let $x_1, \ldots, x_m$ be samples which are independent and identically distributed (*i.i.d*) because they are drawn from the same processes, the goal of statistical inference is then to use the data $x_1, \ldots, x_m$ to estimate the parameters $\overrightarrow{\theta}$ governing the assumed model $L(.)$. Let $P(.)$ be the probability that the parameters $\overrightarrow{\theta}$ govern the assumed distribution, conditional on the observed data/samples $x_1, \ldots, x_m$. Using Bayes' rule, statistical inference can be represented as:

$$\begin{aligned} P(\overrightarrow{\theta} \mid x_1, \ldots, x_m) &= \frac{L(x_1, \ldots, x_m \mid \overrightarrow{\theta}) P(\overrightarrow{\theta})}{D(x_1, \ldots, x_m)} \\ &= \frac{L(x_1, \ldots, x_m \mid \overrightarrow{\theta}) P(\overrightarrow{\theta})}{\int_{\Theta} L(x_1, \ldots, x_m \mid \overrightarrow{\theta}) P(\overrightarrow{\theta}) d\overrightarrow{\theta}} \end{aligned} \tag{2.1}$$

for $\overrightarrow{\theta} \in \Theta$ where $D(.)$ is simply a function of the observed data $\overrightarrow{x}$, and $P(.)$ is a probability distribution of the parameters.

There are two general approaches to statistical inference and parameter estimation, the *frequentist* and *Bayesian* approaches. They are different in their source of variability, the mathematical objects involved, and in the estimation and inference processes. Their difference lies in the source of randomness, where Bayesian uses random variables while the frequentist uses a point estimate.

Under the frequentist approach, the parameter $\overrightarrow{\theta}$ is a constant which makes $P(.)$ insignificant. Hence, $P(.)$ can be absorbed, along with the denominator $D(.)$ from the Bayes' rule

above, into the proportionality sign Aldrich (1997) as

$$P(\vec{\theta} \mid x_1, \ldots, x_m) \propto L(x_1, \ldots, x_m \mid \vec{\theta}).$$

In the frequentist approach, $P(\vec{\theta} \mid x_1, \ldots, x_m)$ is not a probability density, but rather a density of $\vec{\theta}$ where it conveys the relative likelihood of drawing the sample observations given some value of $\vec{\theta}$.

Different from the frequentist approach, in the Bayesian setting, the parameters are latent random variables. This means there is one distribution attached to the parameters. Typically, there is some initial guess about the distribution of $\vec{\theta}$ incorporated through the *prior distribution* $P(\vec{\theta})$. The focus of the Bayesian approach is rather estimating the *posterior distribution* of the parameter, i.e. estimating distribution $P(\vec{\theta} \mid x_1, \ldots, x_m)$.

The posterior distribution of $\vec{\theta}$, $P(\vec{\theta} \mid x_1, \ldots, x_m)$, is the distribution of the parameter conditional upon the observed data $x_1, \ldots, x_m$ and provides some sense of uncertainty regarding the estimate for $\vec{\theta}$. Although the true value of $\vec{\theta}$ is unknown, one may suggest where $\vec{\theta}$ may lie in the parameter space. This information can then be incorporated through the prior $P(\vec{\theta})$ in the Bayesian inference. At last, given the information on how the data $x_1, \ldots, x_m$ are generated by $L(x_1, \ldots, x_m \mid \vec{\theta})$ and the prior $P(\vec{\theta})$, the posterior distribution of the parameters can be calculated.

The data generation $x_1, \ldots, x_m$ can be modeled as a joint density function conditional on the estimated parameters of the assumed model. That is,

$$L(x_1, \ldots, x_m \mid \vec{\theta}) = \prod_{i=1}^{m} L(x_i \mid \vec{\theta})$$

In the Bayesian framework, $L(x_1, \ldots, x_m \mid \vec{\theta})$ is called the *likelihood distribution* of the data $x_1, \ldots, x_m$ (as opposed to the likelihood function of $\vec{\theta}$ on the frequentist framework).

Because of the difference between two approaches in the inference procedure, the frequentists and Bayesians differ in the estimation procedure as well. Under the frequentist framework, the likelihood theory of inference is used where the *maximum likelihood*

estimators (MLE) is the single point summary of the likelihood curve. It is the point which maximizes the likelihood function. Under the Bayesian framework, the posterior distribution of $\vec{\theta}$ is the focus of the approach where estimators such as the posterior expectation (or the posterior mean) or *maximum a posteriori* (MAP) estimator are used to summarize the posterior distribution.

2.1.1 Parameter Estimation

Assume that a sample of n i.i.d observations $x_1, \ldots, x_n$, each of which come from some probability distribution governed by an unknown parameter θ (as shown before, θ might be a vector $\theta = (\theta_1, \ldots, \theta_m)$): $L(x \mid \theta)$, where θ belongs to a parameter space Θ. Parameter estimation aims to find the best estimator of θ. Specifically here, two parameter estimators are of interest: the maximum likelihood estimate (MLE) representing the frequentist approach, and the maximum a posteriori (MAP) representing the Bayesian approach. Both of these estimators are to determine the value of θ that is most likely to have drawn $x_1, \ldots, x_n$.

The maximum likelihood estimate (MLE) is one of some systematic ways for parameter estimation in the frequentist framework. In the frequentist framework, the likelihood function $P(\theta \mid x_1, \ldots, x_n)$ is not considered as a probability distribution of θ conditional upon the data, which is obtained via Bayes' rule, but rather a function of the parameters given the data. Hence, the likelihood function is proportional to the joint probability of the data conditional on the parameter:

$$P(\theta \mid x_1, \ldots, x_n) \propto L(x_1, \ldots, x_n \mid \theta)$$

The maximum likelihood estimate of θ is the value of θ in parameter space Θ that maximizes the likelihood function. Mathematically, an MLE is defined as

Definition 2.1.2. *Let $X_1, \ldots, X_n$ be random samples from a distribution with a parameter θ, and $P(.)$ be a likelihood function of θ. Given that $X_1 = x_1, \ldots, X_n = x_n$ have been observed, a*

maximum likelihood estimate of θ*, denoted by* θ_{MLE}*, is a value that satisfies*

$$\arg\max_{\theta \in \Theta} P(\theta \mid x_1, \ldots, x_n)$$

with $P(\theta \mid x_1, \ldots, x_n) \propto L(x_1, \ldots, x_n \mid \theta)$ *and* $L(x_1, \ldots, x_n \mid \theta) = \prod_{i=1}^{n} L(x_i \mid \theta)$.

For many models, the MLE can be found analytically as an explicit function of the observed data. This involves taking the first derivative of the likelihood function, setting it to 0, and solving the parameter θ. For other models, however, no closed-form solution to the maximization problem is known or available, and an MLE can only be found via numerical optimization. Several optimisation algorithms such as gradient descent, or simulated annealing can be used to find an MLE.

As opposed to the frequentist framework, in the Bayesian framework, $P(\theta \mid x_1, \ldots, x_n)$ is considered as a probability distribution of θ conditional upon the data $x_1, \ldots, x_n$. $P(\theta \mid x_1, \ldots, x_n)$ is now called the posterior distribution. Given the data, all information about the parameter θ can be extracted from $P(\theta \mid x_1, \ldots, x_n)$. Therefore, the posterior distribution can be used to find points or interval estimates of θ. One way to obtain a point estimate is to choose the value of θ that maximizes the posterior distribution $P(\theta \mid x_1, \ldots, x_n)$. This is called the maximum a posteriori (MAP) estimation. Although a pure Bayesian approach does not take single point estimates but rather a full posterior distribution, one could argue that the MAP estimate works on a posterior distribution, not only the likelihood, which comes up in Bayesian setting.

The maximum a posteriori of θ is the value of θ in parameter space Θ that maximizes the posterior distribution. Formally,

Definition 2.1.3. *Let* $X_1, \ldots, X_n$ *be a random sample from a distribution with a parameter* θ*,* $P(\theta \mid X_1, \ldots, X_n)$ *be a posterior distribution of* θ*, and* $P(\theta)$ *be a prior distribution of* θ*. Given that* $X_1 = x_1, \ldots, X_n = x_n$ *have been observed, a maximum a posteriori of* θ*, denoted by* θ_{MAP}*, is a value that satisfies*

$$\arg\max_{\theta \in \Theta} P(\theta \mid x_1, \ldots, x_n)$$

with $P(\theta \mid x_1,\ldots,x_n) \propto L(x_1,\ldots,x_n \mid \theta)\ P(\theta)$, *and* $L(x_1,\ldots,x_n \mid \theta) = \prod_{i=1}^{n} L(x_i \mid \theta)$.

Similar to the MLE, the MAP estimate can be calculated either analytically or numerically. An analytic calculation can be done when the mode(s) of the posterior distribution can be given in a closed form. This is the case when conjugate priors are used. Otherwise, numerical optimization such as the conjugate gradient method or Newton's method should be used. This usually requires first or second derivatives, which have to be evaluated analytically or numerically. From the point of view of Bayesian inference, the MLE is a special case of the MAP estimate which assumes a uniform prior distribution of the parameters. Therefore, the MAP estimate can be seen as a regularization of the MLE.

2.2 Discrete Probability Distribution

This section is dedicated to providing the basic concepts of probability distributions which are either used in, or closely related to, the thesis. Some distributions such as Poisson distribution and negative binomial distributions are described in length due to their use in this thesis.

2.2.1 Bernoulli Distribution

One of the simplest probability distributions for a discrete random variable is a *Bernoulli distribution*. A random variable which follows a Bernoulli distribution is a random variable that can only take two possible values, usually 0 and 1. This random variable models random experiments that have two possible outcomes sometimes referred to as "success" and "failure". An example would be a coin toss where the outcome is either heads or tails.

Definition 2.2.2. *A random variable* X *is a Bernoulli random variable with parameter* p*, i.e.* $X \sim Ber(x \mid p)$*, if its probability mass function is given by*

$$Ber(x \mid p) = \begin{cases} p & \text{if } x = 1 \\ 1-p & \text{if } x = 0 \end{cases}$$

where $0 \leq p \leq 1$ *and* $x \in \{0, \ldots, 1\}$.

2.2.3 Binomial Distribution

A binomial distribution can be thought of as simply the probability of a "success" or "failure" outcome in an experiment that is repeated multiple times. The binomial distribution is frequently used to model the number of successes x in a sample of size n drawn with replacement from a population $\mathcal{N}$.

Definition 2.2.4. *A random variable X is a binomial random variable with parameter n and p, i.e. $X \sim B(x \mid n, p)$, if its probability mass function is given by*

$$B(x \mid n, p) = \binom{n}{x} p^x (1-p)^{n-x}$$

where $0 \leq p \leq 1$ *and* $x \in \{0, \ldots, n\}$.

For a single trial, i.e., $n = 1$, the binomial distribution is reduced to a Bernoulli distribution. Moreover, a binomial distribution $B(x \mid n, p)$ can be obtained by n independent Bernoulli distributions $Ber(y \mid p)$. It is proven that a binomial random variable $B(x \mid n, p)$ is a sum of n independent Bernoulli random variables $Ber(y \mid p)$.

2.2.5 Poisson Distribution

The Poisson distribution is one of the most widely used probability distributions. It is usually used in scenarios where counting the occurrences of certain events in an interval of time or space is of interest. The Poisson situation can also be used to model scenarios of rare events where the distribution typically acts as a replacement of the Binomial distribution. The probability distribution of a Poisson variable is defined as

Definition 2.2.6. *A random variable X is a Poisson random variable with parameter λ, i.e.*

$X \sim Poi(x \mid \lambda)$, *if its probability mass function is given by*

$$Poi(x \mid \lambda) = \frac{e^{-\lambda}\lambda^{x}}{x!}$$

with $\lambda > 0$ *and* $x \in \{0, \ldots, n\}$.

The Poisson distribution holds unique properties where the expected value of the random variable X is equal to the variance of it which is λ. Given a series of samples $x_1, \ldots, x_n$ generated by $Poi(x \mid \lambda)$, and each sample is i.i.d, the joint probability distribution for $x_1, \ldots, x_n$ is the product of Poisson pmfs:

$$\begin{aligned} Poi(x_1, \ldots, x_n \mid \lambda) &= \prod_{i=1}^{n} Poi(x_i \mid \lambda) \\ &= \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!} \\ &= \frac{\lambda^{\sum_{i=1}^{n} x_i} e^{-n\lambda}}{\prod_{i=1}^{n} x_i!} \end{aligned}$$

Recall that in parameter estimation, the likelihood function is set to be proportional to the joint probability of the underlying distribution. Therefore, the likelihood function of λ is defined as

$$L(\lambda \mid x_1, \ldots, x_n) \propto \lambda^{\sum_{i=1}^{n} x_i} e^{-n\lambda}$$

then the MLE of λ can be found by

$$\arg\max_{\lambda \in \mathbb{R}^+} \lambda^{\sum_{i=1}^{n} x_i} e^{-n\lambda}$$

which results in

$$\lambda_{MLE} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

with $\mathbb{R}^+$ being the positive real numbers, and λ_{MLE} being the MLE of λ.

Within the Bayesian framework, the parameter λ is estimated via Bayes' rule

$$\text{posterior} \propto \text{likelihood} * \text{prior}$$

Also, recall that a conjugate prior distribution gives rise to a posterior distribution of similar functional form (Interested readers are encouraged to read Fink (1997) for more detailed explanation about conjugate priors). The conjugate prior of the Poisson distribution is the gamma distribution which is defined as follows:

Definition 2.2.7. *A continuous random variable X is said to have a gamma distribution with parameters $\alpha > 0$, and $\beta > 0$, i.e. $X \sim Gam(\lambda \mid \alpha, \beta)$, if its probability density function is given by*

$$Gam(\lambda \mid \alpha, \beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda}$$

where $\Gamma(\alpha)$ is the Gamma function ($(\alpha-1)!$ for integers and $\int_0^{\infty} x^{\alpha-1} e^{-x} dx$ for non-integers), $x > 0$.

Parameters α, β set the overall shape of the Gamma distribution. An easy way to loosely interpret the prior parameters α, β (assuming they are integers) is that prior to seeing the current samples, we have seen β 'occurrences', $x'_1, \ldots, x'_{\beta}$, with the total cumulative count $\alpha = x'_1 + \ldots + x'_{\beta}$. Since the gamma distribution is the conjugate prior of Poisson distribution, the posterior distribution of the Poisson likelihood and a Gamma prior is:

$$\begin{aligned} P(\lambda \mid x_1, \ldots, x_n) &\propto Poi(x_1, \ldots, x_n \mid \lambda)\, Gam(\lambda \mid \alpha, \beta) \\ &\propto \left(e^{-n\lambda} \lambda^{\sum_{i=1}^{n} x_i}\right) * \left(\frac{\beta^{\alpha}}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda}\right) \\ &\propto \lambda^{\sum_{i=1}^{n} x_i + \alpha - 1} e^{-\lambda(n+\beta)} \\ &= Gam\left(\lambda \mid \sum_{i=1}^{n} x_i + \alpha, n + \beta\right) \end{aligned} \tag{2.2}$$

which involves adding the sample counts $\sum_{i=1}^{n} x_i$ to the (hyper)-parameter α of the gamma prior and adding the number of observations n to the (hyper)-parameter β of the gamma prior distribution.

The Poisson distribution can be seen as the limit of the binomial distribution. This theorem is called *Poisson Limit Theorem*. Suppose $X \sim B(x \mid n, p)$ where n is a very large number and p is a very small positive value. Specifically, assume that $\lambda = n * p$ is a positive

constant. It is proven that the probability mass function of X can be approximated by the probability mass function of a $Poi(x \mid \lambda)$. This is a good theorem since Poisson PMF is much easier to compute than the binomial.

Definition 2.2.8. *Let $X \sim B(x \mid n, p = \frac{\lambda}{n})$, where $\lambda > 0$ is fixed. The Poisson limit theorem states that for any $x \in \{0, 1, 2, \ldots\}$, the following*

$$\begin{aligned} \lim_{n\to\infty} B_X(x \mid n, p) &= \frac{e^{-\lambda}\lambda^x}{x!} \\ &= Poi_X(x \mid \lambda) \end{aligned}$$

holds.

2.2.9 Negative Binomial Distribution

The negative binomial distribution relates to the random experiment of repeated independent binary trials until m successes have been observed. The distribution can be thought of a joint distribution of binomial distribution and Bernoulli distribution. Formally, a negative binomial distribution is defined as follows.

Definition 2.2.10. *A random variable X is a negative binomial random variable with parameter m and p, i.e. $X \sim NB(x \mid m, p)$, if its probability mass function is given by*

$$NB(x \mid m, p) = \begin{pmatrix} x-1 \\ m-1 \end{pmatrix} p^m (1-p)^{x-m}$$

where $0 \leq p \leq 1$ and $n \geq m$.

Interestingly, the negative binomial distribution can be viewed as a generalization of the Poisson distribution. It is a mixture of a family of Poisson distributions with gamma mixing weights. The negative binomial distribution can be viewed as a Poisson distribution where the Poisson parameter is itself a random variable, distributed according to a gamma distribution. Thus the negative binomial distribution is also known as a Poisson-Gamma mixture. To prove

this, let X be a random variable drawn from a Poisson distribution with parameter λ, i.e.

$$Poi(X = x \mid \lambda) = \frac{e^{-\lambda}\lambda^{x}}{x!},$$

and let the arrival rate λ be also modelled as a random variable to represent its uncertainty. The λ is modelled according to a gamma distribution with shape α, and scale β, i.e.

$$Gam(\lambda \mid \alpha, \beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)}\lambda^{\alpha-1}e^{-\beta\lambda}$$

Then, the unconditional distribution of X is obtained by summing out λ in

$$\begin{aligned}
P(X = x) &= \int_{\lambda=0}^{\infty} P(x \mid \lambda)\, P(\lambda)\, d\lambda \\
&= \int_{\lambda=0}^{\infty} Poi(x \mid \lambda)\, Gam(\lambda \mid \alpha, \beta)\, d\lambda \\
&= \int_{\lambda=0}^{\infty} \frac{e^{-\lambda}\lambda^{x}}{x!}\, \frac{\beta^{\alpha}}{\Gamma(\alpha)}\lambda^{\alpha-1}e^{-\beta\lambda} \\
&= \frac{\beta^{\alpha}}{x!\,\Gamma(\alpha)}\, \frac{\Gamma(x+\alpha)}{(\beta+1)^{x+\alpha}} \\
&= \binom{x+\alpha-1}{x}\left(\frac{\beta}{\beta+1}\right)^{\alpha}\left(\frac{1}{\beta+1}\right)^{x} \\
&= NB\left(x \mid \alpha, \frac{\beta}{\beta+1}\right)
\end{aligned}$$

which follows a negative binomial distribution with parameter $m = \alpha$, and $p = \frac{\beta}{\beta+1}$.

The characteristic of a negative binomial distribution is that the variance is greater than the mean. Whereas, in a Poisson distribution, the variance is set to be equal to the mean of the distribution. Thus, the unconditional distribution of $N(t_i, t_j)$ is more dispersed than its conditional distributions since the unconditional distribution follows a negative binomial distribution. The uncertainty in the parameter variable λ has the effect of increasing the unconditional variance of the mixture distribution of $N(t_i, t_j)$.

2.2.11 Multinomial Distribution

The multinomial distribution is a general version of the binomial distribution. It is used to model experiments where there are more than two outcomes, i.e. modelling the probability of counts for rolling an m-sided die in n times. Recall that the binomial random variable has the following properties:

- Fixed number of n trials.
- Each trial is an independent event.
- Only two outcomes are possible (Success and Failure).
- Probability of success (p) for each trial is constant.

A multinomial experiment is almost identical with one main difference: a binomial experiment can have two outcomes, while a multinomial experiment can have multiple outcomes. If the binomial distribution models outcome of n experiments, where the outcome of each trial has a Bernoulli distribution, then multinomial distribution models outcome of n experiments, where the outcome of each trial has a categorical distribution. Formally, a multinomial distribution is defined as

Definition 2.2.12. *Let $\vec{x}$ be a vector of $x_1, \ldots, x_m$, and $\vec{p}$ be a vector of $p_1, \ldots, p_m$. A random variable X is a multinomial random variable with parameter n and $\vec{p}$, i.e. $X \sim Multi(\vec{x} \mid n, \vec{p})$, if its probability mass function is given by*

$$Multi(\vec{x} \mid n, \vec{p}) = \frac{n!}{x_1!, \ldots, x_m!} p_1^{x_1} \times \ldots \times p_m^{x_m}, \quad \text{for } x_i = 0, 1, \ldots \text{ and } \sum_{i=1}^{m} x_i = n$$

where $0 \leq p_i \leq 1$ *for* $i = 1, \ldots, m$ *and* $\sum_{i=1}^{m} p_i = 1$.

2.3 Poisson Processes

The Poisson process is one of the most widely-used counting processes. It is generally used in situations where counting the number of events that occur at a certain rate, but completely at random is the main objective. Since all contents in this thesis are closely related to the Poisson process, a proper definition of counting processes is necessary for readers to understand the Poisson process better.

Definition 2.3.1. *A random process $N(t_i, t_j), t_i, t_j \in [0,\infty)$ is said to be a counting process if $N(t_i, t_j)$ is the number of events occurred from time t_i up to and including time t_j. For a counting process, the following holds:*

1. *$t_i \leq t_j$, for all $t_i, t_j \in [0,\infty)$;*
2. *$N(t_i, t_i) = 0$;*
3. *$N(t_i, t_j) \in \{0, 1, 2, \ldots\}$;*
4. *for $t_i \leq t_s \leq t_j$, $N(t_i, t_j) - N(t_i, t_s)$ shows the number of events that occur in the interval $(t_s, t_j]$.*

For example, you might have a random process $N(t_i, t_j)$ that shows the number of customers who arrive at a supermarket by time t_j (i.e., afternoon) starting from time t_i (i.e., morning).

Definition 2.3.2. *Let $X(t_i, t_j), t_i, t_j \in [0,\infty)$ be a continuous-time random process. $X(t_i, t_j)$ has independent increments if, for all $t_i \leq t_1 \leq t_2 \leq \ldots \leq t_j$, the random variables*

$$X(t_i, t_2) - X(t_i, t_1), X(t_i, t_3) - X(t_i, t_2), \ldots, X(t_i, t_j) - X(t_i, t_{j-1})$$

are independent.

Recall that for a counting process, $N(t_i, t_j) - N(t_i, t_{j-1})$ is the number of events in the interval $(t_{j-1}, t_j]$. Therefore, a counting process has independent increments if the number of

events in non-overlapping intervals

$$(t_i, t_1], (t_1, t_2], \ldots, (t_{j-1}, t_j]$$

are independent. With the definition of independent increments in a counting process, then calculating a joint distribution of two counting processes becomes a simple multiplication of distributions from each process.

Definition 2.3.3. *Let $X(t_i, t_j), t_i, t_j \in [0, \infty)$ be a continuous-time random process. $X(t_i, t_j)$ has stationary increments if, for all $t_i \leq t_1 < t_2 \leq t_j$, and all $r > 0$, the two random variables $X(t_i, t_2) - X(t_i, t_1)$ and $X(t_i, t_2 + r) - X(t_i, t_1 + r)$ have the same distributions.*

The stationary increments can also be stated that the distribution of the difference depends only on the length of the interval, and not on the exact location of the interval on the real line. With those three definitions, a Poisson process can now be properly defined.

Definition 2.3.4. *Let $\lambda > 0$ be fixed. The counting process $N(t_i, t_j), t_i, t_j \in [0, \infty)$ is called a Poisson process with arrival rates λ if all the following conditions hold:*

1. $N(t_i, t_i) = 0$*;*
2. $N(t_i, t_j)$ *has independent and stationary increments;*
3. *the number of arrivals in any interval of length $\tau > 0$ follows the Poisson distribution $Poi(x \mid \lambda\tau)$.*

A Poisson process can also be defined as a pure birth process where in an infinitesimal time interval Δ there may occur only one event. This happens with the probability $\lambda\Delta$ independent of events outside the interval. Let $N(t_i, t_j)$ be a Poisson process with rate λ. Consider a very short interval of length Δ. Then the number of events in interval Δ has the same distribution as $N(t_i, t_{i+\Delta})$. In particular, this can be written

$$\begin{aligned} Poi(N(t_i, t_{i+\Delta}) = 0 \mid \lambda\Delta) &= e^{-\lambda\Delta} \\ &= 1 - \lambda\Delta + \tfrac{\lambda^2}{2}\Delta^2 - \ldots \text{ (Taylor Series)} \end{aligned}$$

Note that if Δ is small, the terms that include second or higher powers of Δ are negligible compared to Δ. Hence, the above equation can be re-written as:

$$Poi(N(t_i, t_{i+\Delta}) = 0 \mid \lambda\Delta) = 1 - \lambda\Delta + o(\Delta) \tag{2.3}$$

with $o(\Delta)$ as a function that is negligible compared to Δ as $\Delta \to 0$, i.e. $\lim_{\Delta\to 0} \frac{o(\Delta)}{\Delta} = 0$.

Equation 2.3 is defined for $N(t_i, t_{i+\Delta}) = 0$. In the case of $N(t_i, t_{i+\Delta}) = 1$, the equation is redefined into:

$$\begin{aligned} Poi(N(t_i, t_{i+\Delta}) = 1 \mid \lambda\Delta) &= e^{-\lambda\Delta}\lambda\Delta \\ &= \lambda\Delta\Big(1 - \lambda\Delta + \tfrac{\lambda^2}{2}\Delta^2 - \ldots\Big) \text{ (Taylor Series)} \\ &= \lambda\Delta + \Big(-\lambda^2\Delta^2 + \tfrac{\lambda^3}{2}\Delta^3 \ldots\Big) \\ &= \lambda\Delta + o(\Delta) \end{aligned} \tag{2.4}$$

Similarly, $N(t_i, t_{i+\Delta}) \geq 2$ can be shown as:

$$Poi(N(t_i, t_{i+\Delta}) \geq 2 \mid \lambda\Delta) = o(\Delta) \tag{2.5}$$

As a matter of fact, all the equations above provide a way to formally redefine a Poisson process.

Definition 2.3.5. *Let $\lambda > 0$ be fixed. The counting process $N(t_i, t_j), t_i, t_j \in [0, \infty)$ is called a Poisson process with arrival rates λ if all the following conditions hold:*

1. $N(t_i, t_i) = 0$;
2. $N(t_i, t_j)$ *has independent and stationary increments;*
3. *Equation 2.3, 2.4, and 2.5 are met.*

It is often useful to consider a more general type of process in which the arrival rate varies as a function of time. Let $N(t_i, t_j)$ be the number of customers arriving at a fast food

restaurant between time t_i, i.e. morning, and t_j, i.e. afternoon. The Poisson process is suitable for this scenario. However, it is quite unlikely that the process has a constant rate λ over time. For example, it is typical to have higher arrival rates of costumers during lunch time compared to, say around 3 p.m. to 4 p.m. In similar scenarios like this, a nonhomogeneous Poisson process might be suited better than the standard (homogeneous) Poisson process. Such a process has all the properties of a Poisson process, except for the fact that its rate is a function of time, i.e., $\lambda = \lambda(t_i, t_j)$.

Definition 2.3.6. *Let $\lambda(t_i, t_j) : [0,\infty) \times [0,\infty) \rightarrow [0,\infty)$ be an integrable function. The counting process $N(t_i, t_j), t_i, t_j \in [0,\infty)$ is called a nonhomogeneous Poisson process with arrival rates $\lambda(t_i, t_j)$ if all the following conditions hold:*

1. $N(t_i, t_i) = 0$;

2. $N(t_i, t_j)$ *has independent increments;*

3. *for any $t_i, t_j \in [0,\infty)$, the following criteria:*

$$\begin{aligned} Poi(N(t_i, t_{j+\Delta}) - N(t_i, t_j) = 0 \mid \lambda(t_i, t_j)\Delta) &= 1 - \lambda(t_i, t_j)\Delta + o(\Delta) \\ Poi(N(t_i, t_{j+\Delta}) - N(t_i, t_j) = 1 \mid \lambda(t_i, t_j)\Delta) &= \lambda(t_i, t_j)\Delta + o(\Delta) \\ Poi(N(t_i, t_{j+\Delta}) - N(t_i, t_j) \geq 2 \mid \lambda(t_i, t_j)\Delta) &= o(\Delta) \end{aligned}$$

are met.

In a nonhomogeneous Poisson process with rate $\lambda(t_i, t_j)$, the number of arrivals in any interval $(t_{i'}, t_{j'})$ is a Poisson random variable; however, its parameter can depend on the location of the interval. Formally, it is written as:

$$N(t_i, t_{j+s}) - N(t_i, t_j) \sim Poi\left(x \mid \int_{t_j}^{t_{j+s}} \lambda(t_j, t_{j+t})dt\right)$$

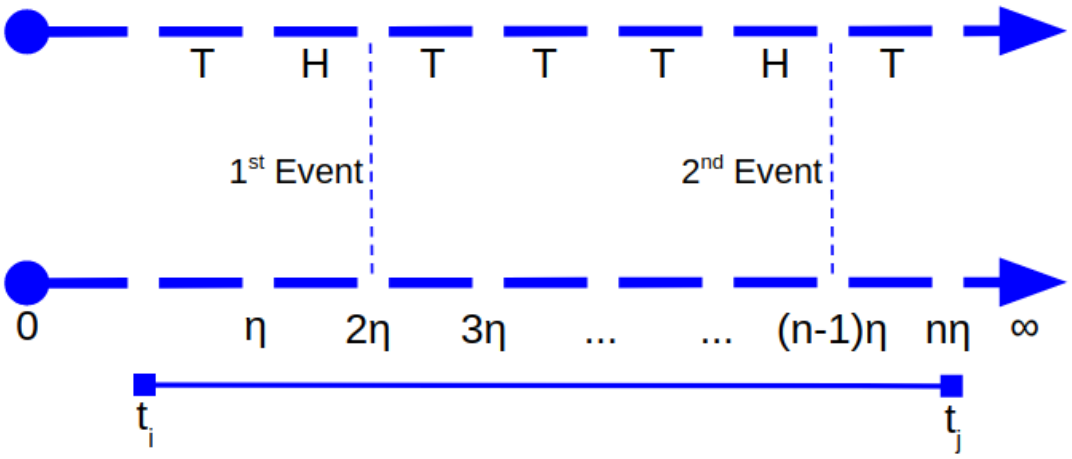

Figure 2.1: An illustrative example of a Poisson process as a limit of a Bernoulli process.

2.3.7 Poisson Process as the Limit of a Bernoulli Process

Assume that the definition of Poisson process has not been made and one would like to model the arrival of events at a rate λ per unit time. At time $t_j = t_i = 0$, no event has been registered, so $N(t_i, t_j) = 0$. The interval $[0, \infty)$ is now divided into small subintervals of length η such that each subinterval corresponds to a time slot of length η. Therefore, each subinterval has a time slot in the form of $\big((k-1)\eta, k\eta\big]$. Let's assume that in each time slot, a coin is tossed for which the probability of the head occurs is $P(H) = p = \lambda\eta$. If the head occurs in a toss, it is assumed that there is an arrival in that subinterval. Otherwise, there is no arrival in that subinterval. Figure 2.1 illustrates this process. If a head event occurs at k-th coin toss, an arrival is said happened at time $t = k\eta$.

Let $N(t_i, t_j)$ be the number of heads that appear from time i to time j. If there are $n \approx \frac{(t_j - t_i)}{\eta}$ time slots in the interval $(t_i, t_j]$, $N(t_i, t_j)$ is the number of heads in n coin tosses. It can be concluded that $N(t_i, t_j)$ follows a binomial distribution with parameters n and p, i.e. $B(N(t_i, t_j) \mid n, p)$. Recall that p is the probability of head in each coin toss, and the value of p is equal to $\lambda\eta$, so

$$\begin{aligned} n \times p &= n \times \lambda \times \eta \\ &= \frac{(t_j - t_i)}{\eta} \times \lambda \times \eta \\ &= \lambda \times (t_j - t_i) \end{aligned}$$

Using the Poisson limit theorem as $\eta \to 0$, the density of $N(t_i, t_j)$ converges to a Poisson

distribution with rate $\lambda \times (t_j - t_i)$. In other words, the number of head events in any interval of length $\tau = (t_j - t_i)$ follows a Poisson distribution $Poi(N(t_i, t_j) \mid \lambda\tau)$ as $\eta \to 0$.

If several non-overlapping intervals are considered, the number of heads events in each interval is solely determined by the results of the coin toss for that particular interval. Since each coin toss is independent from one to another, the aforementioned counting process has independent increments. Thus, all of these meet the definition of a Poisson process informally.

Part I

Statistical Inference from Periodic and Incomplete Data

CHAPTER 3

RELATED WORK ON PARAMETER ESTIMATION IN COUNT DATA

This chapter provides a self-contained review of previous work on estimation techniques for time series data in general and specifically for count data. The chapter focuses on work that is related to the first objective of this thesis, which is modelling regular patterns in a long time series. The chapter starts with general well-known techniques which are typically used to model time series data. Some specific approaches, that have similar goals to this thesis, are explained at length. Finally, other techniques are presented which might be useful for developing a temporal counting model to be used by an agent with limited perceptual abilities.

3.1 Modulated Poisson Process

The Markov modulated Poisson process (MMPP) is typically used in queuing theory. There is not much statistical literature referring to this process. The MMPP is termed, by some literature, a partially observable Poisson process, where observable Poisson arrival rates are

regulated by unobservable states of the Markov processes Ludkovski & Sezer (2012), Meier-Hellstern (1987), RydÃl'n (1996), Prabhu & Zhu (1989), Davison & Ramesh (1996). Ludkovski and Sezer Ludkovski & Sezer (2012) employ an MMPP to solve a decision-making problem. The aim is to maximise the decision maker's expected reward by finding an optimal regime policy for unobservable regimes, which are states of an underlying Markov processes. Davison and Ramesh Davison & Ramesh (1996) apply an MMPP to a binary time series of precipitation data by numerically optimizing the MMPP likelihood.

The MMPP arises from the fact that sample counts of a Poisson process can be contaminated by an outside source. Suppose that true counts x_0 in an interval $[t_i, t_j)$ for $i, j \in 0, \ldots, T$ follow a Poisson process $Poi_0(x \mid \lambda_0)$. A contamination source present for the whole interval $[t_i, t_j)$ produces contaminating counts x_1 according to an independent Poisson process $Poi_1(x \mid \lambda_1)$. All the counts are recorded by $\tau = (\tau_i, \ldots, \tau_{j-1})$ in the fixed interval $[t_i, t_j)$, such that an event at time $t, t > i$ is the $(t-i)$-th event of the count in interval $[t_i, t_j)$. The contamination process follows a Markov process with transition matrix

$$\Gamma = \begin{bmatrix} c_{00} & c_{01} \\ c_{10} & c_{11} \end{bmatrix}$$

Let $C(t) = 1$ if a contamination is present, and $C(t) = 0$ otherwise with $t \in [t_i, t_j)$ so that $C(.)$ is a random step function of continuous time. Notice that $C(t) = 1$ is not enough to imply that the event at time t is contaminated, such that the count for interval $[t_i, t_j)$ is also contaminated. Let $y = (y_{t_i}, \ldots, y_{t_{j-1}})$, where $y_t = 0$ if the event at time t was produced by $Poi_0(x \mid \lambda_0)$, and $y_j = 1$ otherwise. Taken together, $C(.), y, \tau$ convey enough information to recreate counts produced by $Poi_0(x \mid \lambda_0)$ and $Poi_1(x \mid \lambda_1)$. Figure 3.1, taken from Scott (1998), illustrates how the MMPP model works where the honest process refers to x_0, the contamination process refers to $C(.)$, and the additional fraud process refers to x_1.

The goal of this technique is to simulate draws from the posterior distribution of $\theta = (\lambda_0, \lambda_1, c_{01}, c_{10})$ given an observed count x_i, but not the missing data $C(.)$ and $y(.)$. The missing data, however, include a process occurring in continuous time. A Markov Chain Monte

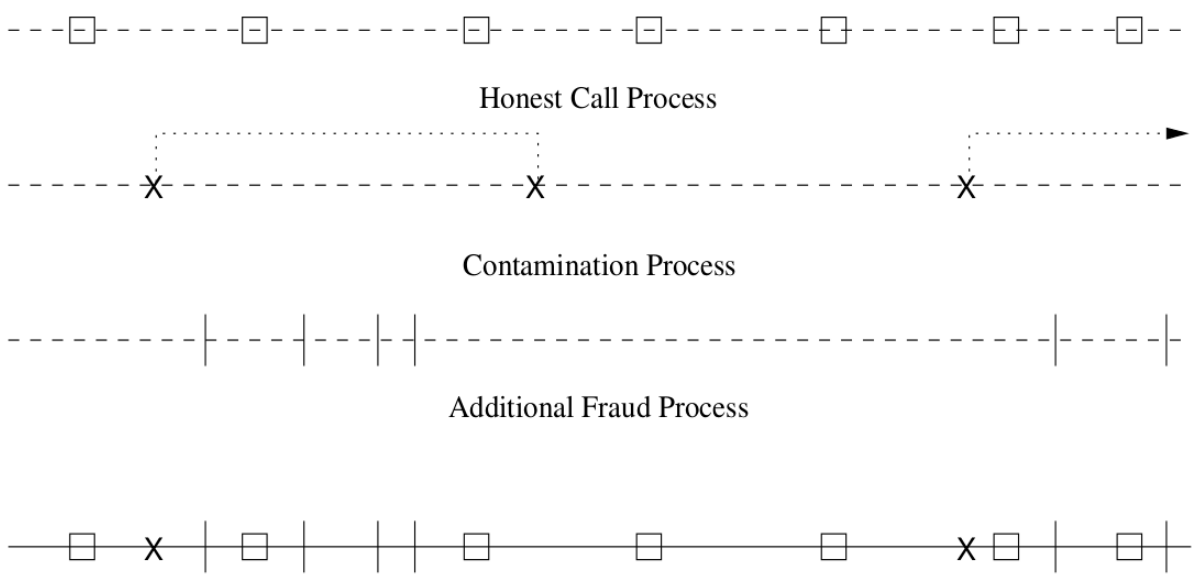

Figure 3.1: Illustration of the MMPP taken from Scott (1998). Notice that the contamination process does not generate contaminated events. The observed process consists of honest events coming from $Poi_0(x \mid \lambda_0)$ and fraudulent events coming from $Poi_1(x \mid \lambda_1)$. However, it is not possible to distinguish honest events from the fraudulent events.

Carlo (MCMC) technique is typically applied to draw the missing data given observed counts $x_1, \ldots, x_n$. Turin Turin (1996) proves that maximum likelihood estimates of MMPP parameters can be obtained by the EM algorithm.

Scott Scott (2001, 1998) modifies an MMPP to a Markov modulated non-homogeneous Poisson process (MMNHPP), which adds an ability to model posterior probabilities of Poisson arrival rates as a function of time. It combines the aspects of the MMPP and the non-homogeneous Poisson process to model point processes, both regular patterns (governed by the non homogeneous Poisson processes) and irregular bursts of activity (governed by the MMPP). This technique is successfully applied in network fraud detection web surfing behaviour and freeway traffic data Scott (2001), Ihler et al. (2006), Scott & Smyth (2003), Ihler et al. (2007). Hutchins et al. Hutchins et al. (2007) extends the work of Ihler et al. (2006) from utilising a single source of information to multiple sources, so as to estimate building occupancy over period of time. The extension could recover missing or corrupted information

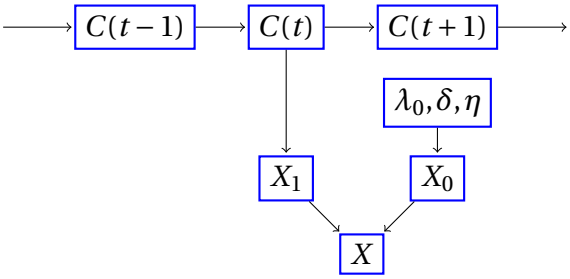

Figure 3.2: Graphical representation of the MMNHPP. The parameters λ_0, δ, η are the periodic components of $\lambda(t_i, t_j)$. The Markov structure of $C(t)$ represents the bursty effect.

in occupancy estimation. Specifically, λ_0 is decomposed into

$$\lambda(t_i, t_j) = \lambda_0 \delta_{d(t)} \eta_{d(t),h(t)},$$

where $d(t)$ ranges from 1 to 7, indicating the day on which time t_i, t_j falls, and $h(t)$ indicates the interval within that day where t_i, t_j falls. One should note that further assumptions are imposed. These are that $\sum_{j=1}^{7} \delta_j = 7$ and $\sum_{i=1}^{D} \eta_{j,i} = D$, where D is the number of time intervals of a day. These assumptions provide an easy interpretation to the values λ_0, δ, η such as λ_0 is the average rate of the Poisson process over a full week, δ_j is the *day effect*, or relative change for day j, and $\eta_{j,i}$ is the relative change in time period t_i, t_j where i lies, given day j (the *time-of-day effect*).

A similar example to MMPP is a non-homogeneous Poisson process where the λ parameter is modulated by Weibull type distribution, which governs how λ should increase (or decrease) over time by small amount. This is used to predict noise exceedances, ozone exceedances, and software reliability issues Guarnaccia et al. (2014), Achcar et al. (2012), Achcar (2001). Specifically, $\lambda(t_i, t_j)$ is transformed into

$$\lambda(t_i, t_j) = \int_{t_i}^{t_j} Wei(t \mid a, r) dt$$

where

$$Wei(t \mid a, r) = \frac{a}{r} \frac{t}{r}^{a-1} e^{-\frac{t}{r}^a}$$

Note that the value of $a < 1$, $a > 1$, $a = 1$ indicates that the rate function $\lambda(t_i, t_j)$ is, respectively,

decreasing, increasing, or constant over time. The vector of parameters of the model to be estimated is $(a, r) \in R^2$, which is typically performed via MCMC. By taking advantage of the natural relationship among posterior and prior distributions of a, r, and the likelihood of the function of the model, the parameters' posterior can be expressed as

$$P(a, r \mid x_1, \ldots, x_n) \propto L(x_1, \ldots, x_n \mid a, r) P(a) P(r)$$

as shown and applied by Guarnaccia et al. (2014), Carlin & Louis (2010) where the unknown parameters (a, r) are sampled using a Gibbs sampling algorithm.

3.2 Spectral Analysis Approach

Spectral analysis is a type of analysis where a stochastic time series can be split into a number of sine and cosine waves with random amplitudes. Spectral analysis deals with a single time series focusing on its frequency components. Its application ranges from astronomical data Templeton (2004), Brault & White (1971), Scargle (1982), to economics Nerlove (1964), signal processing Papoulis & Chamzas (1979), and robotics KrajnÃŋk et al. (2014), Krajnik, Fentanes, Cielniak, Dondrup & Duckett (2014), KrajnÃŋk et al. (2015), Coppola, Krajník, Duckett, Bellotto et al. (2016). The core process of spectral analysis lies in the application of Fourier transformation to extracting frequency components from the time series data. Nerlove successfully showed the advantage of using spectral analysis in identifying seasonal fluctuations in economic time series Nerlove (1964). Among the high frequency components of the time series which are identified by the spectral analysis are those produced by seasonality. Nerlove showed that a slowly changing and stochastic seasonal pattern will reveal itself in the spectrum of an economic time series by a series of peaks at certain frequency Nerlove (1964).

Spectral analysis also shows its benefit in recovering incomplete time series data from its Fourier and inverse Fourier transformation. The application of recovery from incomplete and inaccurate measurement due to noise has been shown both in signal and image processing

CandÃĺs et al. (2006), Stark (1987). Brault and White directly applied spectral analysis, via the fast Fourier transform, to correct solar time series for observational restoration Brault & White (1971). It is based on a filter technique which suppresses the noise and corrects the restoration optimally. Scargle Scargle (1982) modified a classical periodogram technique to provide reliable periodic signal detections which are hidden in noise, with emphasis on the analysis of unevenly spaced data. The modification to a classical periodogram is needed to retain the simple statistical behaviour in the case of unevenly spaced data.

3.2.1 Fourier Analysis

Fourier analysis is the technique of using an infinite number of sine and cosine waves with different frequencies, amplitudes, and phases to represent a given set of data or a function. The process of finding a correct combination of multiple different frequencies, amplitudes and phases is done with a tool called *Fourier transform*. The *Fourier transform* is a reversible, linear transformation that decomposes a function (of time) $f(t)$ into the frequencies $F(\omega)$ that make it up. The Fourier transform of a function $f(t)$ is defined by:

$$\mathscr{F}(f(t)) = F(\omega) = \int_{-\infty}^{\infty} f(t) e^{-2\pi i \omega t} dt \tag{3.1}$$

where ω is the frequency, i is the imaginary square root of -1 used in complex numbers. The function $F(\omega)$ is commonly referred to as the frequency spectrum of $f(t)$. $F(\omega)$ gives how much power $f(t)$ contains at the frequency ω. In addition, f can be obtained from F via the *inverse Fourier transform*:

$$\mathscr{F}^{-1}(F(\omega)) = f(t) = \int_{-\infty}^{\infty} F(\omega) e^{2\pi i \omega t} d\omega \tag{3.2}$$

Equation 3.1 and 3.2 show that the Fourier transform is invertible, and therefore, one can recover the function $f(t)$ from its spectrum $F(\omega)$ and vice versa.

$F(\omega)$ is formed of *complex exponentials*. A complex exponential is a complex number in

the form of

$$e^{i\theta} = \cos(\theta) + i\ \sin(\theta)$$

which is a point on the unit circle at an angle of θ. For any given complex exponential $e^{i\theta} = \cos(x) + i\ \sin(x)$, the *absolute value* is defined as:

$$\mid e^{i\theta} \mid = \sqrt{\cos^2(\theta) + \sin^2(\theta)}$$

such that $\mid e^{i\theta} \mid$ is the distance from the origin to the position of $e^{i\theta}$ in a complex plane, and the *argument* is defined as:

$$\arg e^{i\theta} = \theta.$$

The absolute values and arguments correspond to the amplitudes and phase shifts of the frequency components ω. Complex exponentials are periodic functions, and the set of complex exponentials is complete and orthogonal Bracewell (1965). Moreover, the Fourier transform can represent any piecewise continuous function and minimizes the least-square error between the function and its representation Bracewell (1965).

The Fourier transform is a powerful technique, but it is limited by the amount and quality of data that are transformed. The data limits the benefit of the transformation including the maximum and minimum periodicity, the accuracy of the periodicity, and the minimum statistically significant amplitudes that can be found Templeton (2004). The span of the data also determines the resolution of the Fourier transform, which is the precision to which the frequency may be determined. This means that the longer the data set is, the more precise determination of the frequency will be.

There is a bare maximum of a recommended frequency without introducing error, which is half of the span of the data. This is known as the *Nyquist frequency*. The Nyquist frequency is important not only because it defines the highest frequency (and shortest period) detectable within a given dataset, but also because it defines the maximum sampling rate one needs in order to fully describe variations up to the maximum frequency Templeton (2004).

By applying a Fourier analysis to a set of data, one assumes that everything contained

within a given data is a wave. Thus noise will appear in the Fourier spectrum at some level defined by the strength of the true periodic modulations, relative to the background noise and measurement of this noise level determines the reliability of the results.

3.2.2 Frequency Map Enhancement (FreMEn)

The Frequency Map Enhancement technique (FreMEn) was proposed by Krajnik et al. KrajnÃŋk et al. (2014) for creating spatio-temporal environment representations of long-term robotic deployments. The FreMEn technique is based on Fourier analysis in combination with a Bernoulli distribution to represent the binary state of data. It has been used in many applications, such as in occupancy grids to compress long-term observations Krajnık, Santos, Seemann & Duckett (2014), in topological maps to improve robotic search KrajnÃŋk et al. (2015), Krajnik, Fentanes, Cielniak, Dondrup & Duckett (2014), in path planning Fentanes et al. (2015), and also in predicting human activities Coppola, Krajník, Duckett, Bellotto et al. (2016). The technique can be applied to all models that represent the world as a set of independent components with binary states Krajnik et al. (2015).

FreMEn applies to time series data which represent the data in two distinct states: 0 (absent) or 1 (present). Let $s(t)$ represent the state value at time t, which is affected by a set of periodic processes that can be identified by the Fourier transform, i.e.,

$$F(\omega) = \mathscr{F}(s(t)).$$

Therefore, each of the n most prominent (i.e. of highest absolute value) frequency components ω_k (for $k = 1,\ldots,n$) of the spectrum $F(\omega)$ can be chosen and stored along with its amplitude $abs(\omega_k)$, and its argument $arg(\omega_k)$ in a spectral model $\mathscr{P}$. Then $\mathscr{P}$ can be used to reconstruct the smoothed sequence $\tilde{s}(t)$ of $s(t)$ by means of the inverse Fourier transform, i.e.,

$$\tilde{s}(t) = \mathscr{F}^{-1}(\mathscr{P}).$$

Since the original $s(t)$ only represents binary states (0, 1), the $\tilde{s}(t)$ is transformed further into a probability function

$$p(t) = \begin{cases} 1 & \text{if } \tilde{s}(t) > 1 \\ 0 & \text{if } \tilde{s}(t) < 0 \\ \tilde{s}(t) & \text{otherwise} \end{cases}$$

KrajnÃŋk et al. (2014) suggested to threshold the probability $p(t)$ such that an estimated state $s^0(t)$ of the original state $s(t)$ can be obtained. Using the following equation, the estimated state $s^0(t)$ from the spectral model $\mathscr{P}$, can be used to interpolate or predict the model's state $s(t)$:

$$s^0(t) = p(t) > 0.5$$

It is possible that $s^0(t) \neq s(t)$. To fill the gap between the original function $s(t)$ and the estimated function $s^0(t)$, the time t, whenever the difference between those two functions appears, is stored in an outlier set $\mathscr{O}$. Hence, a full state reconstruction $s(t)$ can be represented by

$$s(t) = s^0(t) \oplus (t \in \mathscr{O}) \tag{3.3}$$

where $\oplus$ represents a binary XOR operation.

The FreMEn model consists of a finite set $\mathscr{P}$ representing the periodic processes and an outlier set $\mathscr{O}$. When a new observation of a real state $s^m(t)$ is obtained at time t, a state estimate $\tilde{s}(t)$ of $s^m(t)$ is calculated using Equation 3.3, and if it differs from $s^m(t)$, then the current time t is added to the set $\mathscr{O}$. As new observations come, the prediction accuracy of $p(t)$ is likely to decline gradually and the set $\mathscr{O}$ is likely to grow.

3.3 Auto Regressive Integrated Moving Average (ARIMA)

The Auto Regressive Integrated Moving Average (ARIMA) model is another class of statistical models which provide analysis and forecasting of time series data. These models aim to describe the autocorrelations in the data. It is a generalization of the simpler Auto Regressive

Moving Average (ARMA) by adding the notion of integration. ARIMA is descriptive as acronym capturing the key aspects of the model itself. They are:

- **AR**: Autoregression. A model that uses the dependent relationship between an observation and some number of previous observations. Formally, an autoregressive model of order p (referred as AR(p) model) can be written as

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \ldots + \phi_p y_{t-p} + e_t,$$

  where c is a constant, e_t is noise, and $\phi_i, 1 \leq i \leq p$ is *lagged parameters*. Changing the parameters results in different time series patterns, whereas changing the e_t (the variance of e_t) will only change the scale of the series, but not the patterns. AR is typically restricted to model stationary data, i.e., time series data whose properties do not depend on the time at which the series is observed Hyndman & Athanasopoulos (2014).

- **MA**: Moving Average. A model that uses past forecast errors in a regression-like model Hyndman & Athanasopoulos (2014). Formally, a moving average model of order q (referred as MA(q) model) can be written as

$$y_t = c + e_t + \theta_1 e_{t-1} + \theta_2 e_{t-2} + \ldots + \theta_q e_{t-q},$$

  where e_t is white noise, and $\theta_i, 1 \leq i \leq q$ is *lagged parameters*. Changing the parameters results in different time series patterns. It is proven that any stationary AR(p) model can be represented as an MA(inf) model Hyndman & Athanasopoulos (2014).

- **I**: Integrated (referred to as parameter d in the ARIMA model). The use of differencing of raw observations (e.g. subtracting an observation from the previous observation) is to achieve a covariance-stationary series.

A complete **ARIMA**(p, d, q) model can be written as

$$y'_t = c + \phi_1 y'_{t-1} + \ldots + \phi_p y'_{t-p} + \theta_1 e_{t-1} + \ldots + \theta_q e_{t-q} + e_t, \tag{3.4}$$

where y'_t is the difference series of order d, i.e. $y'_t = (1-B)^d y_t$ with B as the backshift notation $By_t = y_{t-1}$, and $\phi_i, 1 \leq i \leq p, \theta_j, 1 \leq q$ are the lagged parameters of AR and MA respectively. Then equation 3.4 can be written, using the backshift notation, as

$$\begin{array}{ccccc} (1-\phi_1 B - \ldots - \phi_p B^p) & (1-B)^d y_t & = & c + (1+\theta_1 B + \ldots + \theta_q B^q) e_t \\ \uparrow & \uparrow & & \uparrow \\ \text{AR}(p) & d \text{ differences} & & \text{MA}(q) \end{array}$$

The model order (p, q) of ARIMA models are determined by automatic model selection such as the Aikaike's Information Criterion (AIC) or the Bayesian Information Criterion (BIC) Box et al. (2015). This is typically done after the difference order d is established, since the difference order is to make the given time series data stationary. Once the model order has been identified, the parameters $c, \phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q$ can be estimated by maximum likelihood estimation Hyndman & Athanasopoulos (2014).

ARIMA modellingis one of the most widely-used approaches to time series forecasting. The technique has been applied to forecast commodity prices Weiss (2000), oil Nochai & Nochai (2006), natural gas Buchanan et al. (2001), and electric power markets Contreras et al. (2003). It was empirically pointed out in Schaidnagel & Laux (2013) that the standard ARIMA could not deal well with systematic periodicity (or seasonality).

3.3.1 Multiseasonal ARIMA

Box et al., introduced a concept of Seasonal ARIMA (SARIMA) to accommodate seasonality in time series Box et al. (2015). Williams and Hoel showed that the SARIMA with a one-week seasonal difference successfully modelled univariate traffic condition data streams Williams & Hoel (2003). It outperformed heuristic forecast benchmarks and showed that the statement

in Kirby et al. Kirby et al. (1997) that extending a simple ARIMA model would not have a significant impact on the performance in time series data was incorrect. Tseng, Yu, and Tzeng combined the SARIMA model with the back propagation model to forecast two seasonal time series for data-sets which are small in sample size Tseng et al. (2002). The SARIMA model is used to forecast seasonal time series whose forecasts and residuals are input to the layer of the BP model to minimize the residuals. Given that $d \geq 0$ and $D \geq 0$ and the difference series $y'_t = (1-B)^d \ (1-B^s)^D y_t$ is a stationary ARMA, a seasonal ARIMA$(p,d,q)(P,D,Q)_s$ process with period S is formally defined as

$$\underset{\uparrow}{(1-\phi_1 B - \ldots - \phi_p B^p)} \quad \underset{\uparrow}{(1-\Phi_1 B^s - \ldots - \Phi_P B^{s^P})} y'_t \quad = \underset{\uparrow}{(1+\theta_1 B + \ldots + \theta_q B^q)} \quad \underset{\uparrow}{(1+\Theta_1 B^s + \ldots + \Theta_Q B^{s^Q})} e_t$$

$$\underset{\text{AR}(p)}{} \qquad \underset{\text{SAR}(P)}{} \qquad \underset{\text{MA}(q)}{} \qquad \underset{\text{SMA}(Q)}{} \tag{3.5}$$

with SAR as Seasonal AR, SMA as Seasonal MA.

Box et al., also commented that a multiplicative seasonal ARIMA given by Equation 3.5 can be extended for the case of multiple seasonalities Box et al. (2015). The double seasonal ARIMA was used to model actual measured GSM traces in China with two periodicities of daily and weekly Shu et al. (2003). This model can be expressed as ARIMA$(p,d,q)(P_1,D_1,Q_1)_{s_1}(P_2, D_2,Q_2)_{s_2}$. Laing and Smith set $s_1 = 48$ to model the within-day seasonal cycle of 48 half-hours, and $s_2 = 336$ to model the within-week cycle of 336 half-hours to forecast electricity demand Laing & Smith (1987). The model from Laing & Smith (1987) was then tested and compared to Holt-Winters exponential smoothing formulation by Taylor Taylor (2003). He pointed out that ARIMA modelling is not well-suited for time-series that are dominated by trend and seasonal variation.

When the time-series data are dominated by seasonal variation, Hyndman and Athanasopoulos suggested the use of Fourier series Hyndman & Athanasopoulos (2014) to model multiple seasonality. External regressors, in the form of Fourier terms, are added to an ARIMA

model to account for the seasonal behaviour. The Fourier-ARIMA model is defined as

$$y_t = c + \sum_{i=1}^{M} \sum_{k=1}^{K_i} (\alpha \sin(\frac{2\pi k t}{p_i}) + \beta \cos(\frac{2\pi k t}{p_i})) + y'_t$$

where c as constant (or the drift parameter), y'_t as an ARIMA(p,d,q), $K_i, 1 \leq i \leq M$ as Fourier terms, M as the number of periodicities, and $p_i, 1 \leq i \leq M$ as periodicity. Provided with a set of models, AIC estimates the quality of the model compared to the other models, and thus provides a means for model selection. The value of K_i is chosen so as to minimize the AIC criteria. Tularam and Ilahee applied the spectral-ARIMA to determine whether recent climate variations in Queensland are in anyway unusual in the context of the 50-60 years of rainfall and temperature records Tularam & Ilahee (2010). The 3-year cycle in the rainfall and temperature data set was found easily, suggesting that spectral-ARIMA is an effective method for this variation. The technique was also applied to forecasting daily maximum ambient O_3 concentrations in six European monitoring stations Kumar & De Ridder (2010). Spectral analysis was used to extract the most prominent frequencies from the time series, such that the residuals of this process are stationary and to be modelled using an ARIMA process. An extension to multiple-day-ahead forecasting was proposed with a similar technique Kumar (2015). However, this showed that singular spectrum analysis (SSA) in combination with ARIMA outperformed spectral-ARIMA.

3.4 Gaussian Processes

Gaussian Processes (GPs) are a state-of-the-art non-parametric regression method. GPs have succeeded in performing inference directly over the space of functions Rasmussen (2004). This technique is different than any model which has been discussed so far with a parameterised class of functions and/or a prior over the parameters. In the context of time series modelling, GPs can be used as priors over the functions representing the behaviour of the data. Formally, a GP is a continuous stochastic process where every finite set of random

variables $f = \{f_1, f_2, \ldots\}$, is Gaussian distributed Rasmussen (2004). A GP is fully determined by a mean m, and a kernel/covariance function k. It is denoted as follows

$$(f(x_i), f(x_j), f(x_k), \ldots) \sim \mathcal{N}\left(\begin{bmatrix} m(x_i) \\ m(x_j) \\ m(x_k) \\ \vdots \end{bmatrix}, \begin{bmatrix} k(x_i,x_i) & k(x_i,x_j) & k(x_i,x_k) & \\ k(x_j,x_i) & k(x_j,x_j) & k(x_j,x_k) & \\ k(x_k,x_i) & k(x_k,x_j) & k(x_k,x_k) & \\ & & & \ddots \end{bmatrix}\right)$$

where $m(x)$ and $k(x,x')$ are the mean and covariance functions respectively. The role of the mean function m is straight forward; it specifies the pointwise mean of this distribution over functions. It is common practice to assume zero mean, since marginalizing over an unknown mean function can be equivalently expressed as a zero-mean GP with a new kernel Lloyd et al. (2014). The role of the kernel $k(x,x')$ determines how the model generalizes or extrapolates new data. This represents high level properties of the unknown function, f. Hence, a characteristic of a GP model is defined by specifying a language of kernels. For example, one of the kernels represents periodic functions using covariance of the form

$$k(x_i, x_j) = e^{-\frac{2\sin^2(\pi(x_i - x_j)/p)}{l^2}}$$

where l and p are constants. Whenever $| x_i - x_j |$ is an integer multiple of p the correlation is equal to 1, resulting in exact periodicity of the random functions. Ghassemi and Deisenroth showed the advantage of using GPs and periodic kernel to model data with an underlying periodic structure in it Ghassemi & Deisenroth (2014). They re-parameterized the periodic kernel, which, in combination with a double approximation, allows for analytic long-term forecasting of a periodic state evolution with Gaussian processes. Karunaratne et al., provided novel kernel-combination methods to explicitly model working-week effects in periodic data using Gaussian process regression Karunaratne et al. (2017). They showed that the introduction of hyperpriors in combination with daily and weekly kernels increased the prediction accuracy significantly which, in the end, outperformed ARIMA model.

There are other kernels which are not covered in this section. Brahim-Belhouari and Bermak introduced a nonstationary covariance function for nonstationary time series data Brahim-Belhouari & Bermak (2004). They have successfully shown the effectiveness of the proposed approach with excellent prediction of respiration signals. Wilson and Adams introduced simple closed-form kernels, based on the Fourier transform, in combination with a Gaussian mixture for Gaussian processes to discover patterns and enable extrapolation Wilson & Adams (2013). The simplicity of the proposed kernels showed that the kernels can be used to replace popular kernels, such as the squared exponential kernels. For more detailed information about kernels, Rasmussen and Williams provide constructions of common GP kernels and analyse their properties Williams & Rasmussen (2006).

Many application of GPs are related to time series and forecasting. Girard and Rasmussen presented a novel approach which uses knowledge of the variance on inputs to Gaussian process models to achieve more accurate prediction variance in the case of noisy inputs Girard et al. (2003). Leith et al., showed the superiority of GPs as a potential unifying framework within which many existing forecasting methodologies may be cast, such as Seasonal Auto-Regressive Integrated (SARI) and Basic Structural Models (BSM) in the case of Irish electrical load forecasting Leith et al. (2004).

3.4.1 Automatic Statistician

Automatic Statistician (AS) is a framework used in Lloyd et al. (2014), Duvenaud et al. (2013), Grosse et al. (2012), Lloyd (2015), which employs Gaussian processes as its core model. The framework can automatically represent any time series data with a suitable kernel composition, which may be kernel addition, and/or kernel multiplication.

The essential elements of this framework are a set of base kernels capturing different function properties, and a set of composition rules which combine kernels to yield other valid kernels Lloyd et al. (2014). The base kernels are white noise for statistical noise in the form of a Gaussian distribution where the values at any pair of times are identically distributed and statistically independent, constant for constant functions, linear for linear functions, squared

exponential for smooth functions, and periodic for periodic functions. The composition rules are addition and multiplication:

$$\begin{aligned}(k_1 + k_2)(x, x') &= k_1(x, x') + k_2(x, x') \\ (k_1 \times k_2)(x, x') &= k_1(x, x') \times k_2(x, x')\end{aligned}$$

since any valid kernel (positive semidefinite kernels) is closed under addition and multiplication. Richer structures such as periodic with noise, or periodic with trend can be obtained by combining kernels using the above operations. By summing kernels, data can be modelled as a superposition of independent functions, possibly representing different structures. A sum of kernels can be interpreted as an OR-like operation: two points are considered similar if either kernel has a high value. In time series models, the summation can represent superposition of multiple processes, possibly operating at different scales Duvenaud et al. (2013). By multiplication, interactions between different input dimensions or different notions of similarity can be accounted for. A multiplying kernel is an AND-like operation since two points are considered similar if both kernels have high values. Lloyd et al., extended the aforementioned operations by adding a changepoints operation into the framework to model more realistic time series Lloyd et al. (2014). Changepoints are defined through addition and multiplication with sigmoidal functions:

$$CP(k_1, k_2) = k_1 \times \sigma + k_2 \times \bar{\sigma}$$

where $\sigma = \sigma(x)\sigma(x')$ and $\bar{\sigma} = (1 - \sigma(x))(1 - \bar{\sigma}(x'))$.

Searching over structures given data is done by proposing all base kernel families applied to all input dimensions. The searching procedures follow the set of expressions:

1. Any subexpression $\mathbb{S}$ can be replaced with either $\mathbb{S} + \mathbb{B}$, or $\mathbb{S} \times \mathbb{B}$, where $\mathbb{B}$ is any base kernel family

2. Any base kernel $\mathbb{B}$ may be replaced with any other base kernel family $\mathbb{B}'$.

Duvenaud et al., proposed to search over this space using a greedy search by choosing the highest scoring kernel and expanding it with all possible operators Duvenaud et al. (2013).

The scoring process involves marginal likelihood since it balances the fit and complexity of a model and also, conditioned on kernel parameters, the marginal likelihood of a GP can be computed analytically Rasmussen & Ghahramani (2001).

3.5 Cox Processes and Bayesian Nonparametrics

The frequentist theorist Alen revolutionised counting process model estimation by introducing a general nonparametric statistical theory for the class of multiplicative counting process Aalen (1978). The technique was adapted into the Bayesian framework in the work of Lo for Poisson processes Lo (1982), and Lo and Weng for general multiplicative processes Lo & Weng (1989). Consider a Poisson process with arrival rate λ. Lo showed that a gamma process prior is conjugate. To define the gamma process prior, consider a σ-finite measure, and say the measure μ is selected by a gamma process prior if for disjoint sets $A_1, \ldots, A_k$, the collection of random variables $\{\mu(A_1), \ldots, \mu(A_k)\}$ are independent gamma random variables of scale 1 and means $\alpha(A_i)$. The measure μ is then said to have shape measure α and scale parameter 1. The corresponding probability measure having this finite-dimensional distribution is denoted by $P_{\alpha,1}$. The measure can be rescaled by an α-integrable positive random function β by defining $\beta\mu(A) = \int_A \beta(x)\mu(dx)$ and the corresponding probability measure is denoted $P_{\alpha,\beta}$. Lo showed that if there are independent realisations $N_1, \ldots, N_n$ of N, and we assign a prior measure $P_{\alpha,\beta}$ to the intensity measure λ, the posterior measure is $P_{\alpha+\sum_1^n N_i, \frac{\beta}{1+n\beta}}$.

Consider the special case where $\beta(x) = 1/\theta$, and suppose estimating the arrival rate $\lambda_t = \lambda(0,t) = \lambda(t)$ under the integrated squared error loss is of interest. It is straightforward to verify that the Bayes estimator is

$$\frac{\alpha(0,t)}{\theta+n} + \frac{n}{n+\theta}\frac{1}{n}\sum_{i=1}^{n} N_i(0,t),$$

which is the weighted average of the prior guess and the sample empirical estimate.

In general, the tools to estimate the parameter $\lambda(t)$, assumed integrable over the period

of observation A, of a non-homogeneous Poisson process in a non-parametric way are similar to those for density estimation. Conditional upon the total number of points $N = N(A)$ the points are distributed as order statistics from a distribution with density Cox & Lewis (1966). Diggle used this fact to develop a kernel estimator for the arrival rate Diggle (1985). A type of histogram estimator in setting up a Bayesian analysis has also been used and introduced by Peeling et al. Peeling et al. (2007). However, the work of Diggle opened the creation of a Bayesian structure to assign a prior related to a Gaussian process, typically of the form $e^{X(t)}$ where $X(t)$ is a Gaussian process. By the same misnomer as for the lognormal distribution, this tends to be called a log Gaussian Cox process. Although it is the log intensity which is Gaussian, and serves as a prior for the arrival rate of a non-homogeneous Poisson process, the model mathematically corresponds to a doubly stochastic Poisson process Cox (1955). The conditional likelihood for this model, given the realisation of $\lambda(s), s \in A$, is simply the usual Poisson likelihood

$$L(x \mid \lambda(s)) = e^{\int_A (log\lambda(s) dN_s - \lambda(s) ds)}$$

For random infinite dimensional $\lambda(s)$, the integral in the exponential of the equation above cannot be evaluated explicitly, which makes inference with a prior $Y(t)$ based on a Gaussian process intractable. Cressie and Rathbun Rathbun & Cressie (1994) and Møller et al. MÃÿller et al. (1998) took a discretisation approach to obtain a tractable expression for the likelihood. The idea is to approximate the continuous process $Y(t)$ by a sequence of step functions in the linear case, and values on a grid in the spatial case. It has been shown that the resulting posterior density converges to the true posterior as the discretisation interval shrinks to zero Waagepetersen (2004). Waagepetersen also pointed out the sensitivity of the resulting inference to the discretisation scheme Waagepetersen (2004). It has been shown that this approach can achieve $\mathcal{O}(NlogN)$ performance. However, this approach relies on the choice of discretisation, which makes it not scalable with the dimension of the domain and it suffers from poor scaling with sensitivity Lloyd et al. (2015). Lloyd et al. eliminated the requirement for discretisation, while simultaneously delivering $\mathcal{O}(N)$ performance. The idea is to simply define the intensity function as $\lambda(s) = f^2(x), s \in A$, where f is a Gaussian process achieving a

non-negative prior, and use this in a conditional GP formulation with variational distributions $q(u) = \mathcal{N}(u \mid m, S)$, such that expectations become tractable w.r.t. $q(f)$.

Heikkinen and Arjas used piecewise constant functions with a random number of jumps of random size as a prior on the intensity function Heikkinen & Arjas (2001). However, they treated this not as an approximation to a smooth prior process. It does not follow, for example, that the posterior mean is piecewise constant. In fact, it typically comes out smooth.

$$f(s) = \frac{\lambda(s)}{\int_A \lambda(u)\,du}. \tag{3.6}$$

The representation in equation 3.6 was used by Kottas to develop a different estimation method Kottas (2006). He treated $\gamma = \int_A \lambda(u)\,du$ as a separate parameter and used explicit density estimation tools to estimate $f(s)$. f is estimated as a Dirichlet mixture of scaled beta densities. The Dirichlet process is determined by a precision parameter α, which is given a gamma prior, and a base distribution, which is a function of the location and dispersion of the beta distributions. These are taken to be independent uniform and inverse gamma, respectively.

3.5.1 Cox Processes

The Cox process or doubly stochastic Poisson process is obtained by letting the rate $\lambda(t_i, t_j)$ of the Poisson process vary according to a positive stochastic process, say $\Lambda(t_i, t_j)$. The doubly stochastic Poisson process was introduced by Cox Cox (1955) and later on was named the Cox process by Bartlett Bartlett (1963). It is worth noting here that, except when the rate process is determined by the scientific situation, it is difficult to analyse a doubly stochastic process without having repeated observations, since the model is indistinguishable from a non-homogeneous Poisson process based on a single path Waagepetersen (2004).

Wolpert and Ickstadt modelled a spatial Poisson process with random rate, where the rate measure is a kernel mixture with a gamma measure Wolpert & Ickstadt (1998). The approach was applied in analysing the density and spatial correlation of hickory trees. The

same dataset was also analysed by Møller and Waagepetersen in a Bayesian setting using a non-homogeneous Poisson process with a log-Gaussian prior process Waagepetersen (2004). The Gaussian process has constant mean β, variance σ^2, and an exponential correlation function with range parameter α. The hyperparameters $\beta, \sigma^2, \kappa = log(\alpha)$ need prior distributions as well. For that, Jeffreys priors for the mean and variance, and a uniform prior between-2 and 4 for κ were used. The analysis was highly sensitive to the choice of prior for κ, and the Bayesian method shows a significantly large correlation range. As pointed out above, this problem can also be viewed as a parametric Bayesian analysis of the doubly stochastic Poisson process obtained using a log Gaussian rate function.

A doubly stochastic Poisson process with a gamma process (introduced by Lo in Lo (1982)) being its rate function $\Lambda(t_i, t_j)$ was considered in Gutiérrez-Peña & Nieto-Barajas (2003). This process has parameter function α as the rate function measure and β as the scale process. In the case of constant scale $\beta = b$, the resulting process is what they call a negative binomial process of type 2. To perform a Bayesian analysis, they assigned a gamma process prior to the rate function measure α, and computed a closed form expression for the posterior distribution of α given the data.

Rue et al., in their highly influential paper on integrated nested Laplace approximations (INLA), illustrated how their numerical alternative to Markov chain Monte Carlo methods can be applied to a doubly stochastic Poisson process where the intensity process is log Gaussian Rue et al. (2009). Although the method would work for any positive function of a Gaussian process such that the resulting doubly stochastic Poisson process is valid. The calculation of over 20,000 marginal distributions, applied to the rainforest dataset also analysed by Waagepetersen Waagepetersen (2007), took four hours of computing time. Again, the Gaussian process was discretised to a fine grid. To get similar precision with MCMC methods would be prohibitive computationally. It is possible within INLA to calculate Bayes factors. However, the prior distributions used for the underlying random field are usually improper which means that the Bayes factor is only determined up to an unknown ratio of constants.

3.6 Summary

In this chapter, some selected approaches that have previously been used for modelling temporal and regular data have been reviewed. Classical techniques such as ARIMA have been judged as not well-suited for time series data that are dominated by regular patterns. Widely-used Gaussian processes showed promising results in their application to time series data including the ones with regular patterns in it, but were also found to have weaknesses related to scalability, due to non-analytic dependence upon GPs' parameters. The idea of applying spectral analysis in time series data such as the one in FFT-ARIMA or FreMEn was presented. As illustrated by Krajnik et al. Krajnik, Fentanes, Cielniak, Dondrup & Duckett (2014), spectral analysis with the Fourier transformation is well-suited for time series data with regular patterns. Furthermore, the technique does not suffer from scalability problems.

CHAPTER 4

SPECTRAL POISSON PROCESS

This chapter focuses on building a counting model which is able to sensibly retrieve and represent temporally recurring patterns of count data on extended temporal scales with limited observations. Limited observations come from the fact that the count data were collected by a mobile robot which was patrolling within perimeters of a large area. The robot's inability to be everywhere at any time makes the count of people passing at particular locations temporally sparse.

Bayesian estimators based on time-varying Poisson models in combination with spectral analysis were developed to capture the periodic structures hidden within a long time series of count data. The model, which is called spectral-Poisson model, is able to accommodate the inference of count data at locations with sparse information. Moreover, applied to the scenario of counting the number of people passing, the model is capable to:

1. predict human activity level at particular times, and

2. classify locations based on the periodic patterns of the activity level.

This approach was deliberately chosen as a derivation of the FreMEn approach proposed by Krajnik et al. which has been used in real-world experiments and which successfully models

recurring patterns of binary data KrajnÃŋk et al. (2014), Krajnik, Fentanes, Cielniak, Dondrup & Duckett (2014), KrajnÃŋk et al. (2015), Coppola, Krajník, Duckett, Bellotto et al. (2016). In contrast to this approach, this thesis contributes two extensions:

1. The non-homogeneous Poisson processes are employed as counting states to replace the binary states of FreMEn.

2. A new way of selecting the most prominent frequency components of the Fourier spectrum was proposed.

This chapter starts with how the counting data is represented as a periodic Poisson process. In Section 4.2, the Fourier Transform is briefly re-introduced, and the proposed method to select the most prominent frequency components is explained. The performance of the spectral-Poisson is tested on simulated data, and the result is explained in Section 4.3. Section 4.4 describes the experimental results on real-world data. Section 4.5 studies the best method to cluster spectral processes according to their spectro-temporal signatures. Finally, Section 4.6 concludes the contribution and discusses the limitations of this chapter.

4.1 Counting Representation

The probabilistic counting model is addressed with a Poisson distribution whose the probability mass function is in the form of:

$$Poi(N(t_1, t_2) = x \mid \lambda) = \frac{e^{-\lambda}\lambda^x}{x!} \tag{4.1}$$

where the parameter λ represents *the rate, or average number of occurrences* in a fixed time interval, and $N(t_1, t_2) = x$ is the count over the time interval $[t_i, t_j)$ for $i, j \in \{0, \ldots, T\}$. By transforming λ to be a function of time, i.e. $\lambda(t_i, t_j)$, equation 4.1 becomes a non-homogeneous Poisson distribution, in which the degree of heterogeneity depends on the function $\lambda(t_i, t_j)$.

When $\lambda(t_i, t_j)$ is imposed to be a periodic function of time, then the distribution becomes a periodic Poisson distribution. Formally speaking, a Δ periodic Poisson process is a Poisson process in which a single periodicity Δ is imposed on the function λ such that:

$$\lambda(t_i, t_j) = \lambda(t_{i+\Delta}, t_{j+\Delta}) \tag{4.2}$$

where $i, j \in \{0, \ldots, T\}$. As a fixed time interval δ at any point in time is used, $\lambda(t_i, t_{i+\delta})$ for $i \in \{0, \ldots, T\}$ and $\delta \in \mathbb{N}$ is defined. Counts within t_i, t_{j+s} are sampled as

$$N(t_i, t_{j+s}) - N(t_i, t_j) \sim Poi\left(x \mid \int_{t_j}^{t_{j+s}} \lambda(t_j, t_{j+t}) dt\right)$$

A conjugate prior is chosen to have a fully Bayesian estimation over the parameter $\lambda(t_i, t_j)$ of the Poisson distribution which follows a Gamma density:

$$Gam(\lambda(t_i, t_j) \mid \alpha_{(i,j)}, \beta_{(i,j)}) \quad = \quad \frac{\beta_{(i,j)}^{\alpha_{(i,j)}}}{\Gamma(\alpha_{(i,j)})} \lambda(t_i, t_j)^{\alpha_{(i,j)}-1} e^{-\beta_{(i,j)}\lambda(t_i,t_j)}$$

where $\alpha_{(i,j)}, \beta_{(i,j)}$ are the shape and the rate parameters Fink (1997). The posterior distribution of $\lambda(t_i, t_j)$ given a set of observed counts $x_1, \ldots, x_n$ within particular time interval t_i, t_j is calculated as

$$\begin{aligned} P(\lambda(t_i, t_j) \mid x_1, \ldots, x_n) &\propto Poi(x_1, \ldots, x_n \mid \lambda(t_i, t_j))\, Gam(\lambda(t_i, t_j) \mid \alpha_{(i,j)}, \beta_{(i,j)}) \\ &= Gam\left(\lambda(t_i, t_j) \mid \sum_{i=1}^{n} x_i + \alpha_{(i,j)}, n + \beta_{(i,j)}\right) \end{aligned} \tag{4.3}$$

On each posterior update, the *Maximum a Posteriori (MAP) hypothesis* $\lambda_{map}(t_i, t_j)$ is chosen to be the point estimate of $\lambda(t_i, t_j)$ with

$$\lambda_{map}(t_i, t_j) = \frac{\alpha_{(i,j)} - 1}{\beta_{(i,j)}} \tag{4.4}$$

where $i, j \in \{0, \ldots, \Delta\}$. As a single periodicity Δ is imposed on the parameter $\lambda(t_i, t_j)$ for the periodic Poisson process, $\lambda_{map}(t_i, t_j)$ becomes a periodic function such that a collection of

MAP estimates ordered from $\lambda_{map}(t_0, t_0+\delta)$ to $\lambda_{map}(t_{\delta-1}, t_{\delta-1}+\delta)$ creates a MAP time series.

4.2 Spectral Representation

To capture the periodic structures over the λ function, i.e. $\lambda(t_i, t_j)$ of a periodic Poisson process, the Fourier transformation is proposed which offers a fast transformation and re-transformation. This section is started by defining the mathematical form of the Fourier transformation. The proposed technique to do spectral analysis is then explained in details including the representation of λ function in its spectral form.

4.2.1 Fourier Transform

The *Fourier transform* is a reversible, linear transformation that decomposes a function of time $f(t)$ into the frequencies $F(\omega)$ that make it up. $F(\omega)$ is formed of *complex exponentials*. A complex exponential is a complex number in the form of

$$e^{i\theta} = \cos(\theta) + i\ \sin(\theta)$$

which is a point on the unit circle at an angle of θ. For any given complex exponential $e^{i\theta} = cos(x) + i\ sin(x)$, the *absolute value* is defined as:

$$\mid e^{i\theta} \mid = \sqrt{\cos^2(\theta) + \sin^2(\theta)}$$

such that $\mid e^{i\theta} \mid$ is the distance from the origin to the position of $e^{i\theta}$ in a complex plane, and the *argument* is defined as:

$$\arg e^{i\theta} = \theta.$$

The absolute values and arguments correspond to the amplitudes and phase shifts of the frequency components ω.

Any analysis and/or alteration to the frequency properties of function $f(t)$ can be made

Algorithm 1 *l addition amplitude model (AAM)*

Input: $x_1, \ldots, x_n$: input signal,
total: maximum total frequency
Output: $\mathscr{S}$: a collection of $(\mid \omega_k \mid, arg(\omega_k), \omega_k)$
Procedure:
1. Init. k $\leftarrow$ 1
// Get the frequency zero ($\omega_1 = 0$)
2. $\omega_k \leftarrow DFT(x_1, \ldots, x_n)[0]$
3. $\mathscr{S}[\omega_k] \leftarrow [\mid \omega_k \mid, arg(\omega_k), \omega_k]$
4. Repeat until k > total
   - $k \leftarrow k+1$
   // Get the frequency with the highest amplitude
   - $\omega_k \leftarrow DFT(x_1, \ldots, x_n)[1]$
   // Update $\mathscr{S}$ with ω_k
   - if $\omega_k \in \mathscr{S}$, $\mid \mathscr{S}[\omega_k] \mid \leftarrow \mid \mathscr{S}[\omega_k] \mid + \mid \omega_k \mid$
     $arg(\mathscr{S}[\omega_k]) = avg([arg(\mathscr{S}[\omega_k]), arg(\omega_k)])$
     else $\mathscr{S}[w_k] \leftarrow [\mid \omega_k \mid, arg(\omega_k), \omega_k]$
   // Create a cosine signal from ω_k
   - $x'_1, \ldots, x'_n \leftarrow \mid \omega_k \mid * cos(2\pi * \omega_k + arg(\omega_k))$
   // Subtract current $x_1, \ldots, x_n$ with the cosine signal
   - $x_1, \ldots, x_n \leftarrow x_1, \ldots, x_n - x'_1, \ldots, x'_n$

by altering the frequency spectrum $F(\omega)$ in the frequency domain. Once the alteration is complete, the transformation back to the domain of $f(t)$ is typically carried out to the altered frequency spectrum $F'(\omega)$. This kind of process is referred to as *spectral analysis*.

4.2.2 Spectral Analysis

Recall that we model the counting problem by a periodic Poisson process described in 4.1, the parameter $\lambda(t_i, t_j)$ is represented by its MAP estimate $\lambda_{map}(t_i, t_j)$. Since $\lambda(t_i, t_j)$ is a function of time, its representation MAP estimate $\lambda_{map}(t_i, t_j)$ is also a function of time with which follows Δ periodicity, i.e. $\lambda_{map}(t_i, t_j) = \lambda_{map}(t_i + \Delta, t_j + \Delta)$. For simplicity, the periodicity extraction is applied to the $\lambda_{map}(t_i, t_j)$, i.e. point estimates, rather than to its Gamma distribution. Periodic patterns of $\lambda_{map}(t_i, t_j)$ can be extracted using the Fourier transform by calculating the frequency spectrum $F(\omega)$ of $\lambda_{map}(t_i, t_j)$, i.e. $F(\omega) = \mathscr{F}(\lambda_{map}(t_i, t_j))$. Once $\lambda_{map}(t_i, t_j)$ is in frequency domain, a spectral analysis on $F(\omega)$ can be carried out. One simple

spectral analysis is to select l frequency components ω_k (for $k = 1,\ldots,l$) with the highest absolute value creating a new frequency spectrum $F'(\omega)$. Then the inverse Fourier transformation is performed on $F'(\omega)$ to reconstruct a smooth periodic function of $\lambda_{map}(t_i,t_j)$, i.e. $\lambda'_{map}(t_i,t_j) = \mathscr{F}^{-1}(F'(\omega))$. This spectral analysis, which we call *l Best Amplitude Model (BAM)*, has been adopted by KrajnÃŋk et al. (2014), Krajnik, Fentanes, Cielniak, Dondrup & Duckett (2014), KrajnÃŋk et al. (2015), Coppola, Krajník, Duckett, Bellotto et al. (2016).

A gradual frequency component selection during the spectral analysis on $F(\omega)$ is proposed. To obtain periodic patterns within $\lambda_{map}(t_i,t_j)$, a component ω_k with the highest absolute value is selected, and a periodic function $f_\omega(t_i,t_j)$ from the component ω_k is created. Then a function subtraction of $\lambda_{map}(t_i,t_j)$ from $f_\omega(t_i,t_j)$ is performed. The result of this subtraction is then transformed back to the frequency domain using the Fourier transformation. The selected component ω_k together with its amplitude $\mid \omega_k \mid$, and its argument $arg(\omega_k)$ is then added to a set $\mathscr{S}$. Whenever the same Fourier component ω_k is obtained, the absolute value of the component is summed, and the argument of the component is averaged before it is stored in $\mathscr{S}$. This process repeats until we obtain l desired Fourier components. The l components which are stored in $\mathscr{S}$ consisting of l triples $\mid \omega_k \mid, arg(\omega_k),$ and ω_k describe the amplitudes, the phase shifts and frequencies of the spectral model of $\lambda_{map}(t_i,t_j)$. The details can be seen in Algorithm 1.

The $l-$AAM technique is adopted from Templeton (2004), and has been successfully applied to obtain multi-periodic pulsation from observed stars which have relatively low pulsation magnitudes. As a consequence, the $l-$AAM technique is able to capture the magnitude of each Fourier component from the original signal better than the standard $l-$BAM technique.

For a detailed comparison between $l-$BAM and $l-$AAM, we created a signal formed of 30 different periodic signals and stretched over 10000 time-points. We performed both spectral analysis techniques ($l-$BAM, and $l-$AAM). The signal and its reconstructions can be seen in Figure 4.1. The figure clearly shows how close the $l-$AAM is to the original signal at each point. This shows that $l-$AAM captures the magnitude of the original signal better than

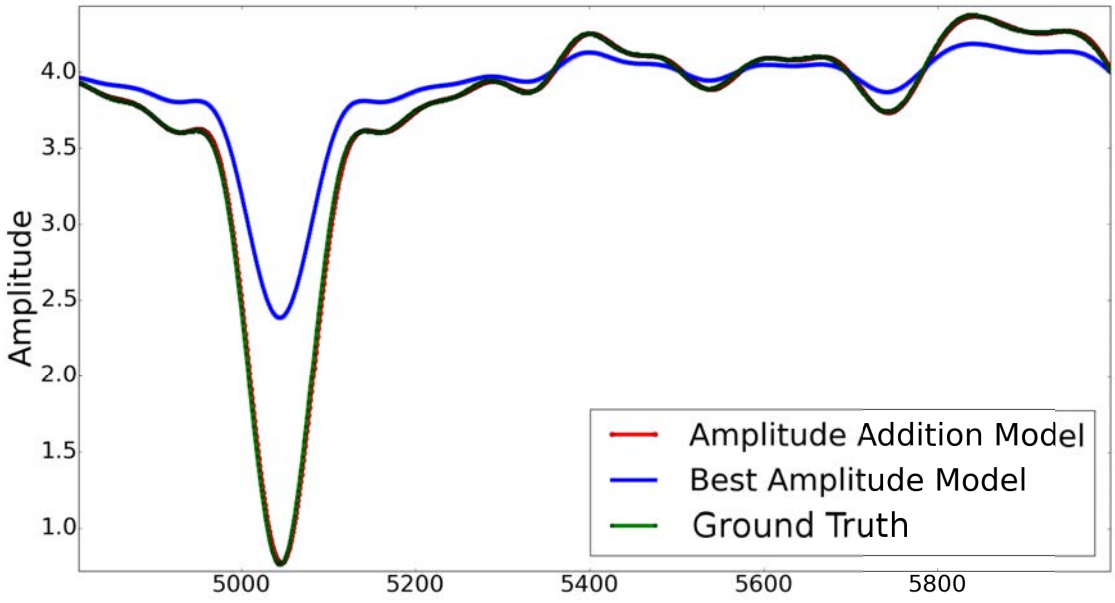

Figure 4.1: A signal composed of 30 different periodic signals and two of its spectral reconstructions (using the l best amplitude model, and the l addition amplitude model).

the $l-$BAM does.

After the spectral analysis has been performed on $F(\omega)$, with either the $l-$BAM or $l-$AAM technique creating a new frequency spectrum $F'(\omega)$, a smoother $\lambda'_{map}(t_i,t_j)$ is constructed via the inverse Fourier transformation, $\lambda'_{map}(t_i,t_j) = \mathscr{F}^{-1}(F'(\omega))$. Given $\lambda'_{map}(t_i,t_j)$ as a function of the new MAP estimates, the corresponding function $\lambda'(t_i,t_j)$ of the periodic Poisson process is then used to replace the original $\lambda(t_i,t_j)$ to achieve a smooth λ function. Replacing the original $\lambda(t_i,t_j)$ is done by reconstructing new Gamma priors $Gam(\lambda'(t_i,t_j) \mid \alpha'_{(t_i,t_j)}, \beta'_{(t_i,t_j)})$ using $\lambda'_{map}(t_i,t_j)$ is where

$$\alpha'_{(t_i,t_j)} = \left(\lambda'_{map}(t_i,t_j) * \beta'_{(i,j)}\right) + 1 \tag{4.5}$$

with $\beta'_{(i,j)} = \beta_{(i,j)}$. This new reconstructed Gamma priors $Gam(\lambda'(t_i,t_j) \mid \alpha'_{(t_i,t_j)}, \beta'_{(t_i,t_j)})$ will be updated once a new sample count x for interval time t_i, t_j arrives and the spectral analysis process explained in this section is repeated.

4.3 Evaluation on Synthetic Data for the Spectral Representation

To investigate the benefit of having spectral analysis via a Fourier transform on the Poisson parameter $\lambda(t_i, t_j)$, the ability of the Fourier transform to recover periodic patterns from synthetic data is evaluated first.

The synthetic dataset is a periodic function $s(t)$ which was built from 30 different periodic signals ω_i with $i \in \{1,\ldots,30\}$, $t \in \{1,\ldots,4\times\Delta\}$, and $\Delta = min(\{\omega_1,\ldots,\omega_{30}\})$. Gaussian noise is then added to each point t in the function $s(t)$. The data is then fed to a periodic Poisson process with the arrival rate $\lambda(t_i, t_j)$ for $i, j \in \{1,\ldots,\Delta\}$. The MAP hypotheses $\lambda_{map}(t_i, t_j)$ are then extracted from the process as point estimates of the periodic Poisson process. These MAP hypotheses $\lambda_{map}(t_i, t_j)$ are then encoded and extracted via spectral analysis either with l-AAM or l-BAM technique to produce a smoother series of point estimates $\lambda'_{map}(t_i, t_j)$ representing l-AAM or l-BAM respectively.

As a comparison to our Fourier reconstruction methods, we include Gaussian processes as a reconstruction technique which is typically used in time series analysis and modelling. Gaussian processes (GP) are a state-of-the-art method for learning models from data. It has been shown that GP with a periodic kernel is able to find repeating patterns for long-term forecasting Ghassemi & Deisenroth (2014). We specifically choose the *Automatic Statistician (AS)* framework of Duvenaud et al. (2013), which employs Gaussian processes. The framework is able to automatically find a suitable kernel composition through kernel addition, kernel subtraction, and/or kernel multiplication. Only five base kernels are chosen to be the kernel composition because the framework are poorly written creating a time-consuming process to construct and calculate the coefficient matrix. $l-$BAM technique is also included in this experiment as a representation of the FreMEn model KrajnÃŋk et al. (2014).

As the synthetic dataset is a periodic function $s(t), t \in \{1,\ldots,4\times\Delta\}$ whose values are fed to the periodic Poisson process, we performed four-fold-cross-validation (CV) on the dataset where each CV-fold is one complete periodic data which has Δ data points. We compared four models including the periodic Poisson processes, $l-$AAM, $l-$BAM, and the Automatic

Table 4.1: Comparison of the predictive accuracy (root mean squared error (RMSE)) of a periodic Poisson process, two spectral reconstructions, and the Automatic Statistician using synthetic datasets.

Method	RMSE	
	no-noise	noise
periodic Poisson processes	101.14	167.98
l Addition amplitude model (AAM)	**101.13**	**161.48**
l Best amplitude model (BAM)	109.11	165.53
Automatic statistician (AS - 5 kernels)	101.57	170.31

Statistician. We recorded the distance of the count data from the point estimates $\lambda_{map}(t_i, t_j)$ for the periodic Poisson processes, or from the smoothened point estimates $\lambda'_{map}(t_i, t_j)$ for the l-AAM, l-BAM, and the Automatic Statistician. The root mean squared error (RMSE) was chosen to be an evaluation metric of the predictive accuracy of the reconstruction and is shown in Table 4.1.

With the absence of noise, Table 4.1 shows that many models performed similarly, with the exception of $l-$BAM which performed poorly. In the presence of noise, $l-$AAM and $l-$BAM performed slightly better than the AS model. The evaluation process on the reconstruction model produced by l-AAM, l-BAM, and the AS model can be thought as a regression test to assure the wanted accuracy of the system after the smoothing procedures.

The AS model might have performed poorly because there is no limitation with what kernel composition it can build to reconstruct the periodic Poisson model even though the prior information is available suggesting periodic kernels. However, even if the kernel composition is restricted to be composed of periodic kernels only, the AS model still requires 30 different periodic kernels representing 30 different periodic signals. Assuming this can be achieved, which was infeasible with limited computational power the author had at that time, the AS model may perform as well as the l-AAM model (This argument is based on the result of the comparison between $l-$AAM and $l-$BAM from Figure 4.1).

One should note that strong uniform Gamma priors for the periodic Poisson processes were used, i.e. $\lambda(t_i, t_j) \approx \lambda(t_{i+1}, t_{j+1})$, which are suitable for later use in our real-world datasets. We did not try to find suitable priors to match our synthetic dataset. As a result, the periodic

Poisson process, which is the basis of all other tested reconstruction models, did not perform well on our synthetic dataset with the average error 100 without noise and 167 with noise.

4.4 Evaluation on Real-World Data

As the previous section shows, the periodic Poisson process with Fourier reconstruction gives a better prediction than a state-of-the-art GP model on a synthetic dataset. The investigation then moved to evaluate the prediction performance of the periodic Poisson process with its Fourier reconstruction techniques on a large, real-world dataset. The prediction performance of the Automatic Statistician is still included as a comparison to the proposed framework. This section starts with a brief explanation of the dataset.

4.4.1 Data Set

The dataset is a collection of *human trajectories* resulting from a long-term deployment of the mobile robot. The data are from a one-month deployment in one of the G4S buildings in Tewkesbury, using a Metralabs Scitos A5 mobile robot equipped with a laser range finder and an RGB-D camera. The robot patrolled on weekdays with a pre-specified schedule (9am - 5pm) and charged at its charging station during weekends and public holidays. A robust human tracking algorithm was used to detect and track humans passing within range of its sensors, whose 2D Cartesian poses were stored in a database Dondrup et al. (2015). Most of these detections, which represent human trajectories, do not represent a complete track of a person. They only represent a small fraction of a person's motion. Moreover, as the robot was patrolling, detections were scattered across places.

Even though approximately 30.000 human trajectories were detected during the deployment, many of them were false positives, including table legs and chairs. To remove false positives, a simple filtering method was used. This is based on the *displacement pose ratio*, which means the distance between the first pose and the last pose of the trajectory over the number of poses in the trajectory Duckworth et al. (2016). A low number of displacement

Figure 4.2: The environment for the dataset with semantic regions annotated. The environment consists of two open plan areas {1,3}, a corridor {12}, a kitchen {5}, three single occupancy offices {2, 10, 11}, three regular offices {4, 8, 9}, and two meeting rooms {6, 7}.

pose ratio for a human trajectory shows that the detected person was stationary. Of 30.000 human trajectories, the 20% of the highest displacement pose ratio were taken as the dataset. With this filtering, false positives still appear, but the number is significantly reduced.

Since the building where our robot was deployed is a large area, we hand-segmented the office into semantic regions such as offices, open plan areas, a kitchen and corridors. From this process, we obtained 12 datasets, one dataset for each semantic region, over a four-week period. The segmented regions can be seen in the global map in Figure 4.2.

Performance on a Real World Dataset

All collected and filtered human trajectories were used as inputs to calculate the posterior $\lambda(t_i, t_j)$ of the periodic Poisson process for each region in the map. As defined in Equation 4.2, a fixed cycle Δ must be chosen to define a periodic Poisson process, hence, Δ was set to a week period for the periodic Poisson process. As a result, each region has four sets of a week worth of trajectories. The MAP point estimates $\lambda_{map}(t_i, t_j)$ are then chosen to represent $\lambda(t_i, t_j)$ for each periodic Poisson process. The $\lambda_{map}(t_i, t_j)$ for $i, j \in \{1, \ldots, \Delta\}$ is then spectro-analysed via Fourier analysis to encode and extract its periodic spectrum $F(\omega)$. The periodic spectrum

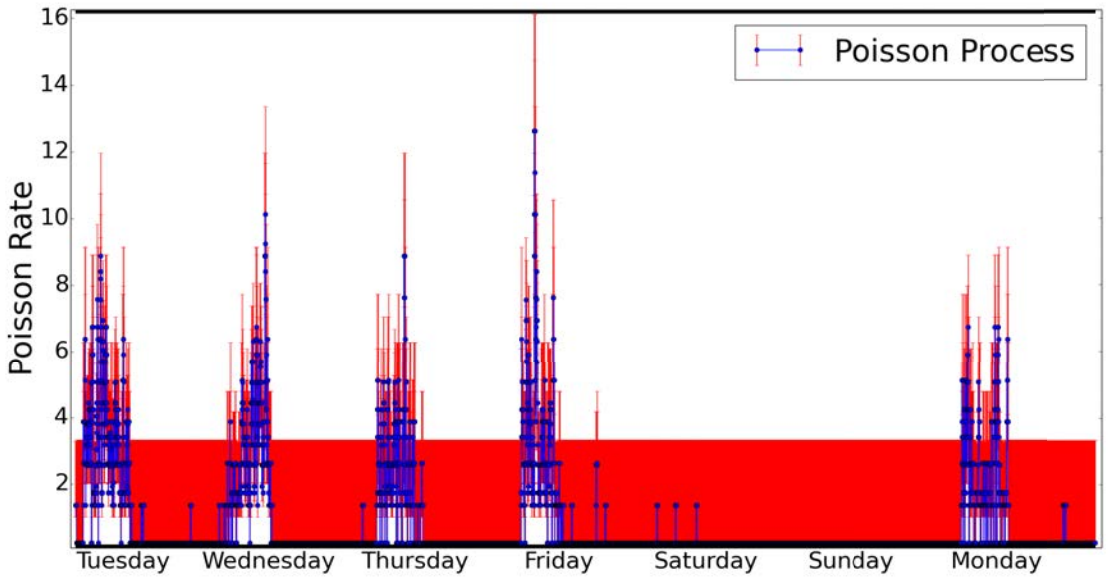

Figure 4.3: The $\lambda_{map}(t_i, t_j)$ of the corridor (region 12) updated over 4 week period.

$F(\omega)$ can be seen as learnt periodicities for $\lambda_{map}(t_i, t_j)$ that represents the region. The learnt periodic spectrums are then used to reconstruct a smooth $\lambda'_{map}(t_i, t_j)$ for better prediction of the level of human activity in a particular space at a particular time. Figure 4.3 shows an example of how the $\lambda_{map}(t_i, t_j)$ of a weekly periodic Poisson process looks after being updated using a four-week long dataset. The red bar at each point in Figure 4.3 shows the upper and lower bound of the confidence interval of each $\lambda(t_i, t_j)$.

We compared the four models described in the previous section in terms of their predictive accuracy on our real-world datasets. We performed four-fold-cross-validation (CV) on the collected dataset where each CV-fold is a different week.

Results are presented in Table 4.2. From the average result, $l-$AAM, $l-$BAM, and the AS model improved the predictive accuracy by 6%, 4.3%, and 3.1% respectively. Once more, the table confirms that in the presence of noise, models with spectral analysis performed slightly better than the one with full Bayesian inference. It can also be concluded from both the synthetic data and real-world data (referring to Tables 4.1 and 4.2) that the introduction of the $l-$AAM technique improves the predictive accuracy.

The time needed for the model reconstruction process can be seen in Table 4.3. In terms

Table 4.2: Comparison of the predictive accuracy of root mean squared error (RMSE) of the periodic Poisson process, two spectral-Poisson models, and the Automatic Statistician using real-world dataset.

Method	RMSE for each region												Avg
	1	2	3	4	5	6	7	8	9	10	11	12	
Poisson	2.6	9.5	7.3	**2.6**	5.8	2.9	4.4	4.1	**1.9**	1.5	4.9	5.0	4.4
$l-$AAM	**2.5**	9.0	**6.6**	**2.6**	**5.7**	**2.4**	**4.2**	**4.0**	**1.9**	**1.4**	**4.8**	**4.7**	**4.1**
$l-$BAM	2.6	9.0	6.9	**2.6**	**5.7**	**2.4**	**4.2**	**4.0**	**1.9**	**1.4**	**4.8**	4.9	4.2
AS	**2.5**	**8.7**	6.7	2.7	5.8	3.4	4.3	**4.0**	2.0	1.5	5.0	4.8	4.3

Table 4.3: Comparison of the computation time of the spectral-Poisson processes and the Automatic Statistician using the real-world dataset. Note that the Automatic Statistician is recorded in hours, while the $l-$AAM and $l-$BAM techniques are in seconds.

Method	Learning time for each region												Avg
	1	2	3	4	5	6	7	8	9	10	11	12	
$l-$AAM	1.1	1.0	1.3	0.8	0.8	0.6	0.8	1.0	0.8	1.0	1.2	1.4	0.9
$l-$BAM	0.1	0.2	0.1	0.1	0.2	0.1	0.1	0.2	0.1	0.3	0.2	0.1	0.2
AS	3.4	1.8	4.0	1.9	2.5	2.4	1.4	2.0	2.0	1.2	2.9	1.5	2.3

of speed, the $l-$BAM reconstruction technique outran other models at least by a factor of 5. The $l-$AAM reconstruction technique is still fast completing in one second on average. This leaves the AS model by far the slowest one, with at least one hour needed to do the reconstruction.

One might argue that the computation time comparison is unfair. Recall that the AS framework has no restriction to what kernel compositions it should build. This created vast possible kernel compositions for the AS framework to test. This resulted in a long time computation time choosing the suitable kernel composition to reconstruct the periodic Poisson process. Restricting the AS framework to look up only on periodic kernels might reduce the computation time significantly. However, the option to pre-choose possible suitable kernels in the AS framework was not available at the time the author evaluated this chapter.

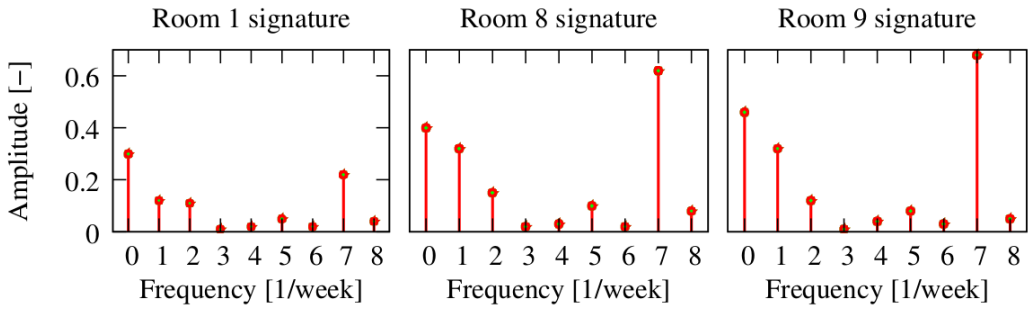

Figure 4.4: The Similarity of the room 8 and 9 signatures indicates that these rooms have a similar pattern of associated arrivals.

4.5 Spectro-Temporal Clustering

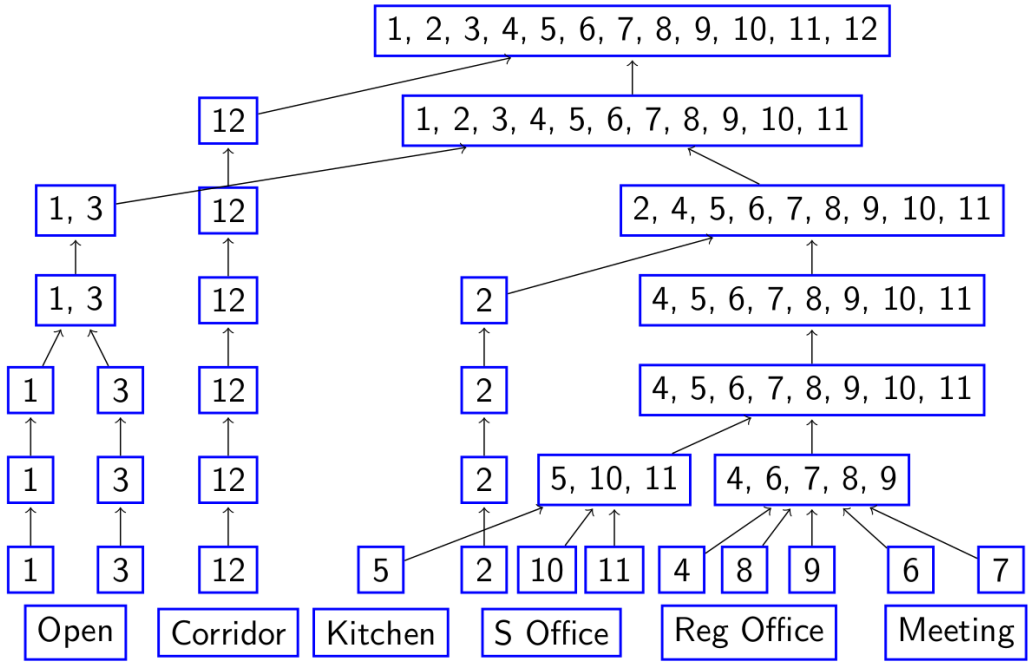

Figure 4.5: DP-Means clustering.

Figure 4.4 shows that some of the regions have a similar frequency spectrum $F(\omega)$. This means that these regions have similar counting patterns. This section is dedicated to testing the aforementioned hypothesis. Moreover, this section also shows whether regions having similar counting patterns are bound to any functionality classification. To do that a clustering process that makes weak prior assumptions about the number of groups of regions and which will produce a hierarchical structure capturing the similarities in the periodicity of the counts across regions is required. For this Dirichlet Process (DP)-means clustering introduced by

Kulis & Jordan (2012) is employed. Inspired by the asymptotic connection between k-means and mixtures of Gaussians, Kulis & Jordan (2012) showed that a Gibbs sampling algorithm for the Dirichlet process mixture approaches a hard clustering algorithm in the limit, and further that the resulting algorithm monotonically minimizes an elegant underlying k-means-like clustering objective that includes a penalty for the number of clusters. Rather than explicitly deciding the number of clusters prior to the learning process, the penalty parameter is varied in using this clustering algorithm. The clustering was compared to the standard K-Means algorithm.

The $l-$AAM reconstruction technique for the periodic Poisson process was used to obtain the frequency spectrum $F(\omega)$ out of the $\lambda_{map}(t_i, t_j)$ for each region. Δ was set to a week period for the periodic Poisson process. The $F(\omega)$, one for each region, was then used as features for the clustering algorithms. Specifically, frequencies extracted by l-AAM construction technique for each region are gathered and they act as features, i.e. columns, for the clustering algorithms. Each region produces a vector of these features where the value on each column represents the amplitude of a frequency corresponding to the column. For example, a region which has frequency 1 (weekly routine) with a value A and frequency 7 (daily routine) with a value B produces a vector in the form of $\{1:A, 2:0, 3:0, 4:0, 5:0, 6:0, 7:B\}$ assuming the common frequencies gathered from all regions are frequency $1, 2, 3, 4, 5, 6, 7$.

Each clustering process constructs a tree which expresses the counting pattern similarities between regions in a hierarchical fashion. For the DP-means clustering, the dendrogram was produced by varying the penalty parameter, whereas, for K-Means, the dendrogram was produced by varying the number of clusters we would like to have. Figures 4.5, and 4.6 show the dendrograms produced by DP-Means clustering and K-Means clustering.

From Figures 4.5 and 4.6, it is easy to verify that the clusters produced by two clustering algorithms are sensible. The clusters formed by the algorithms match roughly with the semantic room type hierarchy. One should note that the single occupancy office $\{2\}$ is a special case. This room belongs to the manager. It thus has quite different periodic arrival patterns to other offices.

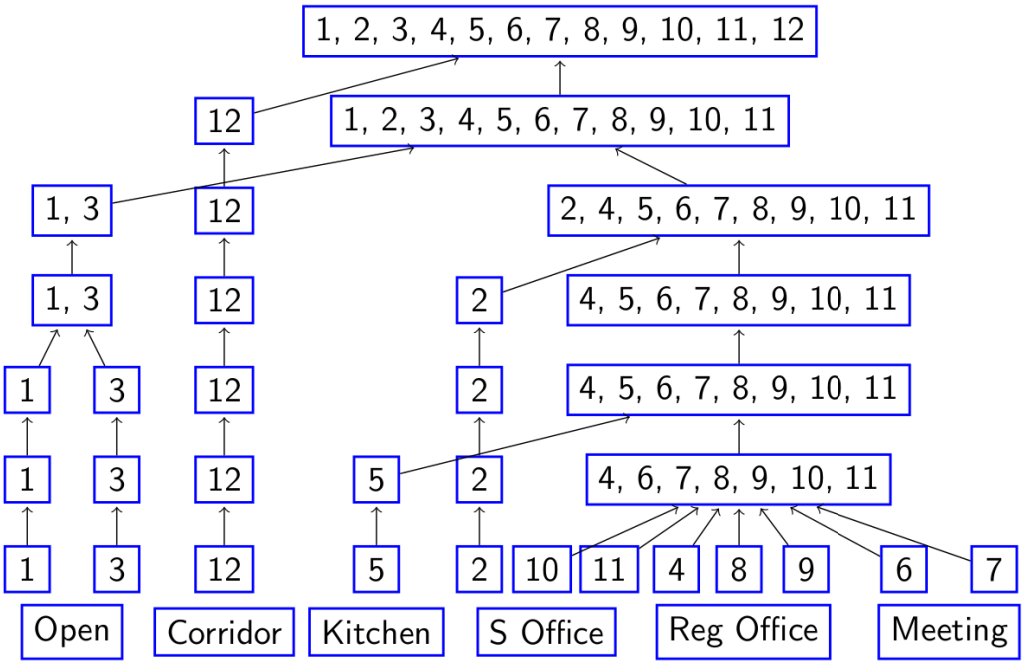

Figure 4.6: K-Means clustering.

4.6 Discussion

This chapter presented a framework to build a probabilistic model of periodic counting processes from observations which are sparse and scattered across time. It is motivated by the fact that the inference made by a counting model built from sparse observations may lead to incorrect conclusions. The framework is based on an assumption that the count data have periodicities which can be observed from the fluctuating number of counts. The count data is modelled by a periodic Poisson process and the periodic patterns are described by means of frequency, amplitude, and phase, modelled using the Fourier and inverse Fourier transforms. By taking the most significant spectral components of the Fourier transform, we indirectly obtain the most significant periodic patterns in the count data.

The application was both on simulated data and real-world data, in counting the level of human activity observed by an autonomous mobile robot. This showed that this framework is able to find regular patterns from sparsely populated observations of count data. It consistently outperformed other temporal modelling methods while being computationally more efficient. The application also shows that each periodic Poisson process has a unique

frequency spectrum which can be further used as features for spectral clustering.

The spectral Poisson process provides a way to deal with representing recurring temporal patterns of count data from temporally incomplete data observations. However, the framework explained in this chapter relies on two things. First, the framework here assumes that the underlying Poisson process has periodicities embedded into it. A Poisson process which is aimed to model trend either locally or globally might be modelled better with other approaches such as Gaussian processes or ARIMA models. Second, the framework also assumes that the event counting algorithms where these counts are made are perfectly reliable.

Part II

Statistical Inference from Unreliable Multi-sensor Data

CHAPTER 5

RELATED WORK ON PARAMETER ESTIMATION WITH SYSTEMATIC ERROR

This chapter specifically aims to provide reviews of previous work on statistical inference with unreliable count data which come from systematic errors. In count data, systematic errors are introduced in each bit, i.e. presence or absence of a person, that makes up the count. Each bit may experience mis-classification and the count resulted from that may end up undercounting or overcounting. The count may also end up having the correct value but the bits making up the count experience mis-classification which counteracts one another. In the review, the focus is on how approaches identify the background or noise and correct the estimation of the parameter of interest.

5.1 Parameter Estimation with Misclassified Counts

The early literature on the effects of misclassified counts involved only binomial and multinomial models Bross (1954), Chen (1979), Hochberg (1977), Tenenbein (1970), Viana et al. (1993). Misclassification happens when there are false positive counts or false negative counts

(or possibly both). False positive counts, which can also be called the overcount, are when the count includes occurrences other than those of interest. Whereas false negative counts, which can also be called the undercount, are when some of the count of interest are missed or omitted. The first technique which was recorded to handle misclassification was double sampling. It was first introduced by Tenenbein to correct for misclassification of binomial data and obtain a maximum likelihood estimate Tenenbein (1970). The double sampling approach utilizes two search techniques to retrieve relevant information: an expensive classification technique to obtain the true count, and a less-expensive classification technique only for error-prone counts. For the expensive classification method, search techniques to obtain the true count, along with the false positive count and false negative count, are used. This part is referred to as a training sample. The training sample typically comes from a small observation-opportunity interval in time or space. In the second, less-expensive, classification method one uses a larger sample with only error-prone counts available. The results of both counts are then combined to obtain estimators for the Poisson rate, and also for the misclassification parameters. The work of Tenenbein was then extended by Chen Chen (1979) and Hochberg Hochberg (1977) to correct mis-classified counts in categorical and multinomial models to obtain maximum likelihood estimates. Viana et al., injected prior distributions and presented Bayesian estimators for the binomial model using the double sampling technique Viana et al. (1993). Bekele extended the work of Viana by introducing a weighted prior scheme and allowing for several sources of information, including expert opinions Bekele (1998).

Moving on from the case of binomial and multinomial models, there are a limited number of works concerning the Poisson model with misclassification. Many deal with the undercount (or under-reporting) problem as this is quite a common problem. Whittemore and Gong estimated cervical cancer rates by taking into account false negative data Whittemore & Gong (1991). Winkelmann and Zimmermann introduced a combination of a Poisson regression model with a logit model for under-reporting, yielding the Poisson-Logistic model. They applied this to model the number of days emloyees were absent from a workplace. The

specification of Poisson regression allows Winkelmann and Zimmermann to relate both the intensity of the Poisson process, as well as the reporting probability, to a set of potential covariates. However, even in a simple model where both the intensity of the Poisson process as well as the reporting probability are assumed constant, the resulting model is not identified based solely on the reported count data Dvorzak & Wagner (2016). Identification requires additional information, which can be provided either by additional data on the reporting process, parameter restrictions or prior information on parameters, e.g., provided by experts. One of the works which illustrated the requirement to have additional information was done by Moreno and Girò Moreno & Girón (1998). They incorporated prior information in a Bayesian analysis of two data sets on assaults. Dvorzak and Wagner assumed a small set of validation data is available, which provides information for the true counts Dvorzak & Wagner (2016). A Bayesian analysis of the Poisson-Logistic model was performed and Bayesian variable selection was incorporated to identify regressors with a non-zero effect and also to restrict parameters of the Poisson-Logistic model.

The effect of misclassification on the Poisson distribution, in the case when the count data may either be undercounting or overcounting, has not been documented as well as the binomial case. Kircher et al. showed that, for a sample of deaths in Connecticut, the cause of death on death certificates often differs from the actual cause, determined by autopsies Kircher et al. (1985). Sposto et al. estimated both cancer and non-cancer death rates, assuming false negatives are possible on both sides of these counts Sposto et al. (1992). The approach used by them followed the frequentist framework. Bratcher and Stamey used a Bayesian method to estimate Poisson rates in the presence of both undercounts and overcounts. They used the double sampling technique where the first sample is searched with both a fallible and an infallible method and the second sample is searched with only a fallible method Bratcher & Stamey (2002). Binomial random variables are used to model the false negative counts, whereas Poisson random variables are used to model the false positive counts. Consider a sample size of Poisson experiments L, and the Poisson rate λ. They assumed the following

models for the various counts

$$\begin{aligned}
&\text{True Count: } && T \sim Poi(t \mid L\lambda) \\
&\text{FN Count: } && (X \mid T = t) \sim B(x \mid t, \theta) \\
&\text{FP Count: } && Y \sim Poi(y \mid L\phi)
\end{aligned}$$

with θ as the rate of false negatives, and ϕ as the rate of false positives. The observed count still follows Poisson as $Z = (T - X) + Y \sim Poi(z \mid L\mu), \mu = \lambda(1-\theta) + \phi$ and the marginal distribution of Z is in the form

$$P(z) = \int_0^1 \int_0^\infty \int_0^\infty f(z \mid \lambda, \phi, \theta) g(\lambda) g(\phi) g(\theta)\, d\lambda\, d\phi\, d\theta$$

with $g(.)$ as prior distribution. The posterior distribution of λ is calculated from $g(\lambda \mid z) = f(z, \lambda)/m(z)$ where $f(z, \lambda)$ may be derived in a similar manner to $P(z)$. The expressions include large sums involving hyper-geometric and confluent hyper-geometric functions. The functions do not have closed form and must be approximated in infinite series form. Because of that, the posteriors are computationally expensive to evaluate. For moderate to large sample sizes, the Gibbs sampler was suggested to approximate these posterior distributions Bratcher & Stamey (2002).

The work of Bratcher and Stamey to estimate the Poisson rate was then extended to a fully Bayesian method for interval prediction of the unobservable actual count in a future sample, given a current double sample Stamey et al. (2004). Two additional Bayesian predictive probability functions for predicting false positives and false negatives were also introduced. Stamey and Young Stamey & Young (2005) managed to obtained closed-form expressions for maximum likelihood estimators (MLEs) of the false negative rate, the false positive rate, and the Poisson rate for the model proposed in Bratcher & Stamey (2002). The estimators are straightforward to calculate and to interpret in terms of evaluating the effectiveness of using unreliable counts.

5.2 Parameter Estimation with Signal and Background Model

Some of the practical problems involving counting processes can be reduced to the case of counting events coming from two independent Poisson-distributed contributions. These two contributions are referred to as "signal" and "background". This approach is quite popular in physics, such as the case of underground detectors which measure the result of high-energy particle collisions. In many papers the aim is to estimate the parameter of interest, although some of them are interested in finding a possible excess which can be attributed to a faint signal. Typically, evidence for the signal is difficult to find and the upper limit of the signal intensity is provided as a replacement Casadei (2011). Moreover, in many practical problems, separating the signal from the background is not an easy job because, usually, only fewer quantities are directly observable, and sometimes only the total number of observed events can be measured Casadei et al. (2017). An attempt to separate the signal from the background has been tried several times in the frequentist framework Bityukov & Krasnikov (2000). The approach using the Bayesian framework was then introduced to tackle similar problems Demortier et al. (2010), Pierini et al. (2011).

To show how signal and background counts are represented, let us assume $x = 0, 1, 2, \ldots$ observed counts following a Poisson distribution. The signal and background sources are independently Poisson distributed with parameters λ and Λ, such that k is sampled from a Poisson distribution with parameter $\lambda + \Lambda : x \sim Poi(x \mid \lambda + \Lambda)$. As many have done, the aim is to estimate the signal rate $\lambda \geq 0$, as is practically always the case in the search for new phenomena Casadei (2012), and treat $\Lambda \geq 0$ as a noise parameter. The result of the statistical inference is provided by the joint posterior probability density

$$P(\lambda, \Lambda \mid x) \propto Poi(x \mid \lambda + \Lambda) P(\lambda, \Lambda)$$

with $P(\lambda, \Lambda)$ as the prior distribution. The reference analysis technique, introduced by Bernado in Bernardo (2005), is the preferred technique since all available information comes

only from the assumed model and the observed data; no additional prior information is available. The core of this technique lies in the formulation of the reference prior $\pi(s,b)$ Berger et al. (2009), defined so as to maximize the amount of missing information, and the result of the inference is the reference posterior obtained by using $\pi(\lambda,\Lambda)$ in place of $p(\lambda,\Lambda)$ in the Bayes theorem above. Assuming one is interested in making inferences about the background rate Λ, an integration of the posterior can be made to obtain the marginal posterior $P(\lambda \mid x) = \int_0^\infty P(\lambda,\Lambda \mid x)d\Lambda$, as the final solution Berger et al. (2009), Bernardo (2005). From $P(\lambda \mid x)$, summary information such as the posterior expectation or MAP hypothesis enclosed by intervals representing some given probability can be computed.

Casadei considered a situation where no prior information about the signal rate λ is available, but the prior estimate of the background expectation $E[\Lambda]$ and standard deviation $\sqrt{V[b]}$ are available Casadei (2012). These two are enough to specify a unique prior density such as a gamma density

$$P(\Lambda) = Gam(\Lambda \mid \alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)}\Lambda^{\alpha-1}e^{-\beta\Lambda}$$

with shape parameter $\alpha > 0$, and scale parameter $\beta > 0$ fixed by requiring $E[\Lambda] = \alpha/\beta$ and $V[\Lambda] = \alpha/\beta^2$. With a prior background following a gamma density, the marginal model becomes

$$\begin{aligned} P(x \mid \lambda) &= \int_0^\infty Poi(x \mid \lambda+\Lambda)Gam(\Lambda)d\Lambda \\ &= \left(\tfrac{\beta}{1+\beta}\right)^\alpha e^{-\lambda} f(\lambda \mid x,\alpha,\beta) \end{aligned}$$

where the polynomial

$$f(\lambda \mid x,\alpha,\beta) = \sum_{m=0}^{x} \binom{\alpha+m-1}{m} \frac{\lambda^{x-m}}{(x-m)!(1+\beta)^m}$$

acts like $(\lambda+\alpha/\beta)^x$ when both α,β are very large.

Hence, the reference prior also depends on the background parameters, and it does so

via Fisher's information function

$$\mid I(\lambda) \mid^{1/2} = \left[\left(\frac{\beta}{1+\beta} \right)^{\alpha} e^{-\lambda} \sum_{n=0}^{\infty} \frac{f(\lambda \mid x, \alpha, \beta)^2}{f(\lambda \mid x+1, \alpha, \beta)} - 1 \right]^{1/2}$$

which involves an infinite sum over terms featuring the polynomial. Since the function is not integrable, Casadei suggested

$$\pi(\lambda) = \frac{\mid I(\lambda) \mid^{1/2}}{\mid I(0) \mid^{1/2}}$$

which makes it trivial to compare it against the uniform prior, so widespread that it can be considered a conventional prior Casadei (2014). When the reference prior $\pi(\lambda)$ been defined, the marginal reference posterior for the signal rate is defined as

$$P(\lambda \mid x) \propto \left(\frac{\beta}{1+\beta} \right)^{\alpha} e^{-\lambda} f(\lambda \mid x, \alpha, \beta) \pi(\lambda)$$

The equation is a proper density where the normalization constant is just the integral of the expression above.

5.3 Summary

This chapter reviewed some approaches which address parameter estimation, especially in the Poisson model, with unreliable count data. The unreliability of the data can be modelled as misclassification or background noise. Techniques which reduce the case of counting events to two independent processes have shown an inability to account counts of interest which are missed or omitted (undercounting). Some other techniques which address the effect of mis-classified counts have addressed mis-classification counts which include undercounts as well as overcounts. However, the posterior is computationally expensive to evaluate and Gibbs sampling was suggested to approximate this posterior distribution.

CHAPTER 6

PARTIALLY OBSERVABLE POISSON PROCESS: INDEPENDENT SENSORS

This and subsequent chapters address the problem of how to efficiently correct miscounts made by multiple unreliable counting processes. Where these counts are made using sensor data, many currently available event counting algorithms have some level of unreliability. This has been shown in Chapter 4 where from 30.000 trajectories collected by a human tracking algorithm, almost 24.000 of them are false positives. This means that large data-sets typically contain systematic errors that lead to bias in the statistical estimates produced by the counting processes.

Practical Bayesian estimators are developed for a partially observable Poisson process (POPP) where miscounts by counters are the only observable information. The work in this thesis deals with a homogeneous Poisson process first. The concept of POPP is introduced where the assumption that miscounts by various counters are conditionally independent from one another given the true count holds.

The term "partially observable Poisson process" sometimes refers to the Markov modulated Poisson process (MMPP) where a Poisson process in which fully observable counts are

conditioned on a latent state which evolves according to a Markov chain Ludkovski & Sezer (2012), Scott (2001), Ihler et al. (2006), Scott (1998), Meier-Hellstern (1987), RydÃl'n (1996), Prabhu & Zhu (1989). The POPP considers the true count as a latent variable. The values of which are inferred from observations made by unreliable counters. In general, this chapter makes the following contributions:

- The POPP poses the problem of how to correctly estimate the parameter of a single homogeneous Poisson process in which true counts are observed through multiple unreliable counters.

- Because the POPP has no conjugate density and its posterior has a number of elements that grow exponentially in the number of observations, the posterior may only be approximated for long sequences of observations. The second contribution of this chapter is to present three tractable approximations to the posterior.

This chapter starts by revisiting the definition of a Poisson process which was introduced in Section 2.3. In Section 6.2, the proposed POPP is described in a mathematical form. Section 6.3 describes all three approximations to the posterior distribution of POPP. The performance of these approximations is evaluated on simulated data, and the results are discussed in Section 6.4. Section 6.5 describes the experimental results on real-world data. This section also showcases the experimental results of partially observable non-homogeneous Poisson processes. Finally, Section 6.6 discusses the limitations of the work presented in this chapter.

6.1 Fully Observable Poisson Process (FOPP)

Let us briefly recall the Poisson process which, from now on, is called the fully observable Poisson process (FOPP).

A fully observable Poisson process is a counting process $N(t_1, t_2)$ where a counter enumerates, without error, the number of events that occurred during a specified interval (t_1, t_2). $N(t_1, t_2) = x_i$ states that in the i-th observation of interval (t_1, t_2), there are x_i events.

The number of events $N(t_1, t_2)$ between t_1 and t_2 follows the Poisson distribution,

$$Poi(N(t_1, t_2) = x \mid \lambda) = \frac{e^{-\lambda}\lambda^x}{x!}$$

where λ represents the *arrival rate, mean count, intensity*, or *expected number of events* in a fixed interval $(t_1,\ t_2)$.

Given a series of events $x_1, \ldots, x_n$ and that the arrival rate λ follows a Gamma prior

$$Gam(\lambda \mid \alpha, \beta) \quad = \quad \frac{\beta^\alpha}{\Gamma(\alpha)}\lambda^{\alpha-1}e^{-\beta\lambda}$$

where α, β are the shape and the rate parameters Fink (1997). Since the gamma distribution is the conjugate density for the Poisson, the posterior for the arrival rate λ is also Gamma:

$$\begin{aligned} P(\lambda \mid x_1, \ldots, x_n) &\propto Poi(x_1, \ldots, x_n \mid \lambda)\ Gam(\lambda \mid \alpha, \beta) \\ &= Gam\left(\lambda \mid \sum_{i=1}^{n} x_i + \alpha, n + \beta\right) \end{aligned} \tag{6.1}$$

6.2 Partially Observable Poisson Process (POPP)

Recall that in the FOPP, there is an assumption that the counter is reliable. This assumption is now removed since many sensor based algorithms have some level of unreliability. Counters may either under-count or over-count. As discussed in Chapter 3, a noise model Dondrup et al. (2015), Schmid et al. (2006) or a filtering algorithm Duckworth et al. (2016), Jovan et al. (2016) might cope well with over-counting. However, under-counting is still an issue. Here the FOPP is altered to incorporate a model of the sensor's reliability to cope both overcount and undercount.

The *true count* (or simply count) and the *sensed count* are distinguished. The true count x_i is the number of events that actually occurred in the i-th sample from the interval (t_1, t_2). Suppose that several sensors are available, and thus the sensed count s_{ji} is the count given by

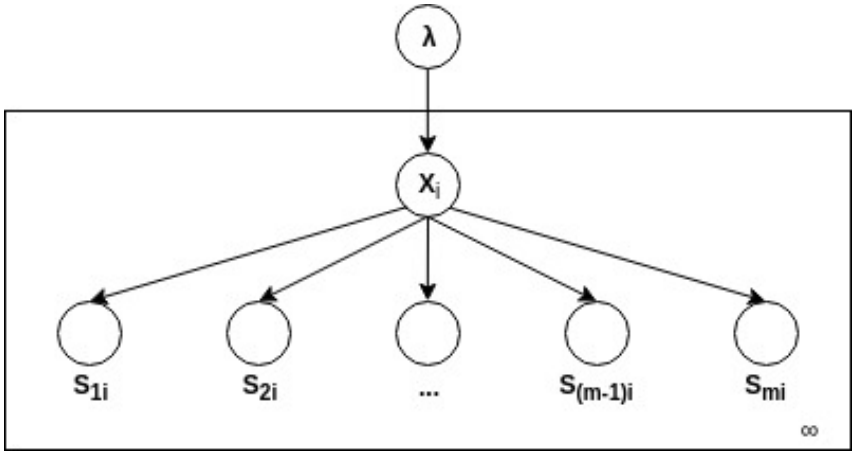

Figure 6.1: Graphical representation of the POPP.

sensor j in the i-th sample from the interval (t_1, t_2). This means that a fully reliable sensor j has $s_{ji} = x_i$ for each sample $1 \leq i \leq n$.

A graphical model with the structure shown in Figure 6.1 is obtained. As shown in the Figure, the true count x_i has become a latent variable which can only be inferred from the sensed count s_{ji} of each sensor j, and the posterior of the Poisson parameter λ is inferred from the posterior of the x_i updated by a vector of m sensed counts $\vec{s_i} = (s_{1i}, \ldots, s_{mi})$.

The parameter λ can be estimated by Bayesian averaging the posterior $P(\lambda \mid x_i)$ over all possible count values x_i with mixing proportions equal to the posterior over x_i given $\vec{s_i}$. The posterior of λ, given n samples $\vec{s} = (\vec{s_1} \ldots \vec{s_n})$, each consisting of m sensors, is:

$$\begin{aligned} P_G(\lambda \mid \vec{s}) &= \sum_{x_1=0}^{\infty} \ldots \sum_{x_n=0}^{\infty} P(\lambda \mid \vec{x})\, P(\vec{x} \mid \vec{s}) \\ &= \sum_{x_1=0}^{\infty} \ldots \sum_{x_n=0}^{\infty} Gam\left(\lambda \mid \sum_{i=1}^{n} x_i + \alpha, n + \beta\right) P(\vec{x} \mid \vec{s}) \end{aligned} \tag{6.2}$$

with $\vec{x} = (x_1, \ldots, x_n)$ for $1 \leq i \leq n$.

$P(\vec{x} \mid \vec{s})$ is a joint posterior over a series of (true) count x_i given the series of sensed count s_i. Factoring $P(\vec{x} \mid \vec{s})$ is started by making the assumption that the vector of sensed counts for sample i $\vec{s_i}$ is conditionally independent of all other vectors of sensed counts for sample j $\vec{s_j}$. Then we assume that *each sensor is conditionally independent of the other sensors given* x_i. Consequently, the probability that the vector of true counts is $\vec{x}$, given n samples of the vector of m sensed counts $\vec{s_1}, \ldots, \vec{s_n}$, is

$$\begin{aligned} P(\overrightarrow{x} \mid \overrightarrow{s_1},\ldots,\overrightarrow{s_n}) &\propto P(\overrightarrow{s_1},\ldots,\overrightarrow{s_n} \mid \overrightarrow{x})\ P(\overrightarrow{x}) \\ &\propto \prod_{i=1}^{n} P(\overrightarrow{s_i} \mid x_i)\ P(x_i) \\ &\propto \prod_{i=1}^{n}\prod_{j=1}^{m} P(s_{ji} \mid x_i)\ P(x_i \mid \overrightarrow{x_{-1}}) \end{aligned} \tag{6.3}$$

where $\overrightarrow{x_{-1}} = x_{i-1},\ldots,x_1$. The independent sensor assumption not only provides convinient calculation for a joint probability distribution of $P(\overrightarrow{x} \mid \overrightarrow{s})$, but also reduces the training data size needed to construct a sensor model for each sensor j (this will be discussed further in Chapter 9, Section 9.2). However, this assumption comes with a drawback which is discussed and solved in Chapter 8.

To complete Equation 6.3, $P(s_{ji} \mid x_i)$ and $P(x_i \mid \overrightarrow{x_{-1}})$ must be defined. The $P(x_i \mid \overrightarrow{x_{-1}})$ can be considered as the conditional distribution of x_i informed by the previous counts $x_{i-1},\ldots,x_1$. It is obtained by imposing a Gamma prior $Gam(\lambda \mid \alpha,\beta)$ on the parameter λ of the Poisson distribution $Poi(x_i \mid \lambda)$ and integrating out λ. It is a Poisson-Gamma mixture which takes the form of a negative binomial distribution.

$$\begin{aligned} P(x_i \mid \overrightarrow{x_{-1}}) &= \int_{\lambda=0}^{\infty} P(x_i \mid \lambda)\ P(\lambda \mid \overrightarrow{x_{-1}})\ d\lambda \\ &= \int_{\lambda=0}^{\infty} Poi(x_i \mid \lambda)\ Gam(\lambda \mid \alpha,\beta)\ d\lambda \\ &= \int_{\lambda=0}^{\infty} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!}\ \frac{\beta^{\alpha}}{\Gamma(\alpha)}\lambda^{\alpha-1}e^{-\beta\lambda} \\ &= \frac{\beta^{\alpha}}{x_i!\,\Gamma(\alpha)}\ \frac{\Gamma(x_i+\alpha)}{(\beta+1)^{x_i+\alpha}} \\ &= \binom{x_i+\alpha-1}{x_i}\left(\frac{\beta}{\beta+1}\right)^{\alpha}\left(\frac{1}{\beta+1}\right)^{x_i} \\ &= NB\left(x_i \mid \alpha, \frac{\beta}{\beta+1}\right) \end{aligned} \tag{6.4}$$

$P(s_{ji} \mid x_i)$ can be thought of as a representation of a noise model for each sensor j. $P(s_{ji} \mid x_i)$ integrates the sensor model of sensor j in calculating how likely sensor j produces a sensed count s_{ji} given the true count x_i. To integrate the sensor model of each sensor j

into the calculation of $P(s_{ji} \mid x_i)$, the Poisson limit theorem, which states that the Poisson distribution may be used as an approximation to the binomial distribution Papoulis & Pillai (2002), is taken advantage of. An arbitrarily close approximation to the probability $P(s_{ji} \mid x_i)$ is defined by assuming there exists a small enough finite subinterval of length δ for which the probability of more than one event occurring is less than some small value ϵ. With this assumption, interval (t_1, t_2) is split into l smaller subintervals $I_1, \ldots, I_l$ of equal size, with the condition that $l > \lambda$ (the condition is crucial since we focus on very small portions of the interval). Consequently, the whole interval $(t_1, t_2) = I_1, \ldots, I_l$ becomes a series of Bernoulli trials, where the k^{th} trial corresponds to whether (1) an event e_k happens with probability λ / l and (2) a sensor j captures the event e_k as the detection d_k at the subinterval I_k.

Given the argument above, the probability of s_{ji} detections given x_i events happened is the aggregate of true positives tp_{ji} in x_i sub-intervals, and false positives fp_{ji} in $l - x_i$ sub-intervals. The probability of a true positive detection (TP) for sensor j in a sub-interval is $tpr_j = P_j(d = 1 \mid e = 1)$, and the probability of a false positive detection (FP) is $fpr_j = P_j(d = 1 \mid e = 0)$. Both tpr_j and fpr_j are the sensor model of sensor j and represent how reliable sensor j is. Thus $P(s_{ji} \mid x_i)$ is defined as a sum of two binomial distributions $B(r \mid n, \pi)$, where the aggregate is constrained to be s_{ji}:

$$P(s_{ji} \mid x_i) = \sum_{\mathrm{tp}_{ji}=0}^{x_i} B\Big(\mathrm{tp}_{ji} \mid x_i, \mathrm{tpr}_j\Big) B\Big(\mathrm{fp}_{ji} \mid \Delta x_i, \mathrm{fpr}_j\Big) \tag{6.5}$$

where $s_{ji} = tp_{ji} + fp_{ji}$, $tpr_j = P_j(d = 1 \mid e = 1)$, $fpr_j = P_j(d = 1 \mid e = 0)$, and $\Delta x_i = (l - x_i)$.

Equation 6.2 now makes the posterior of λ depend not only on the nature of the problem, and the settings of hyper-parameters α and β but also on the reliability of each individual sensor. It also shows the difficulty of belief state estimation in a POPP since there is no conjugate density. Each sensed count $\overrightarrow{s_i}$ sample used to update the posterior of λ adds a factor of countably infinite number of elements. The resulting posterior is a sum of countably infinite sums. One can place an upper bound l on the maximum value of each x_i, but it still makes the number of elements in the posterior grow by a factor l with every sensed count $\overrightarrow{s_i}$.

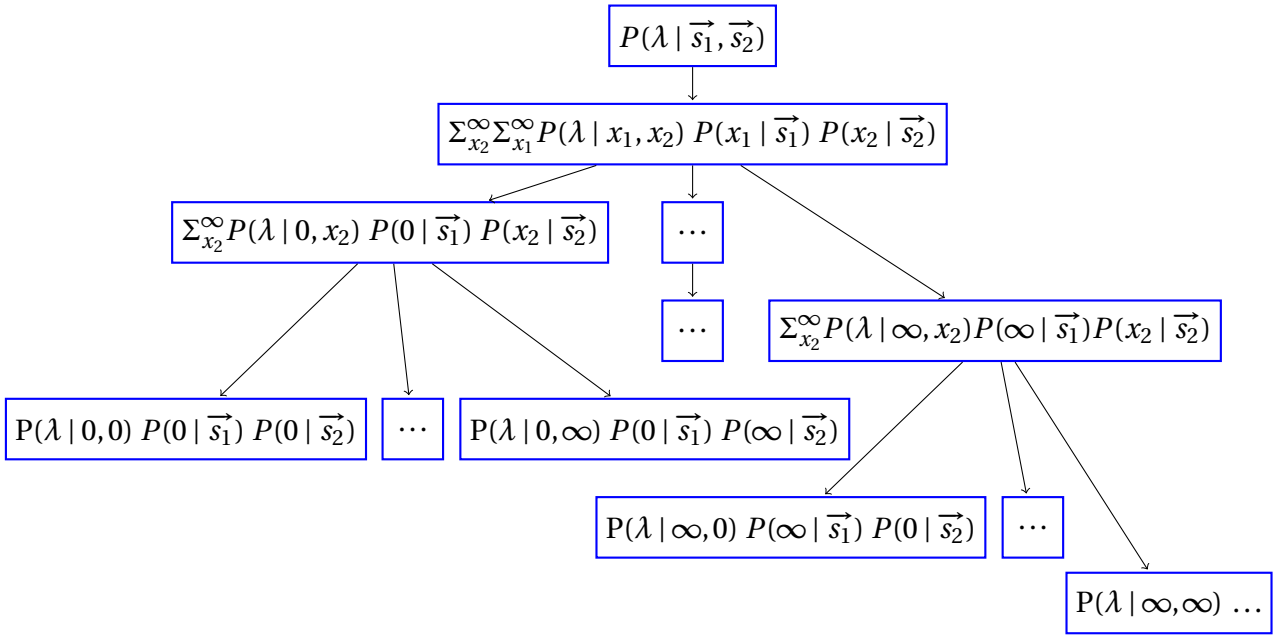

Figure 6.2: Belief state estimation of $P_G(\lambda \mid \overrightarrow{s_1}, \overrightarrow{s_2})$.

For example, assume the upper bound of the true count is l, and two sensed counts $\overrightarrow{s}_1$, $\overrightarrow{s}_2$ are available. Calculating the posterior of λ given the first sensed count $\overrightarrow{s}_1$, i.e. $P_G(\lambda \mid \overrightarrow{s_1})$, produces a sum of l gamma distributions. Calculating the posterior of λ given the second count $\overrightarrow{s}_2$, i.e. $P_G(\lambda \mid \overrightarrow{s_1}, \overrightarrow{s}_2)$, produces a sum of l of a sum of l gamma distributions. This process is depicted by Figure 6.2. As it can be seen from the figure, a parent node is the sum of its children and the number of direct children of a parent is the upper bound l on the maximum value of each x_i.

With this difficulty noted, three different approximations to the posterior are proposed under some reasonable independence assumptions.

6.3 Approximation Filters

To provide an efficient estimator of Equation 6.2, three filters, each of which offers an approximation to the posterior are proposed. First, the posterior Equation 6.2 can be approximated by a single gamma distribution. This approximation gets worse as sensor reliability degrades. A histogram filter is also introduced which is significantly slower in approximating the pos-

terior than the single gamma approximation, but it hovers around it consistently as sensor reliability degrades. Finally, these two approximations can be combined into a switching filter which balances the accuracy against computational cost.

6.3.1 Gamma filter

Algorithm 2 *Gamma filter*

Input: (x_i, y_i): $y_i = P(\lambda = x_i \mid \overrightarrow{s})$
β : $Gam(\lambda \mid \alpha, \beta)$
Output: α, β
Procedure:
// Initialise epoch, error, and maximum step
1. Init. $\eta \leftarrow 0$, $\eta_{max} \leftarrow p$, $err \leftarrow q$, $step \leftarrow r$
// Acquire the mode of $P(\lambda \mid \overrightarrow{s})$
2. $x \leftarrow mode((x_1, y_1) \ldots, (x_m, y_m))$
3. Repeat until $\eta > \eta_{max}$
// Acquire new β' within step boundaries
• $\beta' = random(\beta - step, \beta + step)$
// Acquire α' of Gamma density
• $\alpha' \leftarrow x * \beta' + 1$
// calc $D_{KL}(P(\lambda \mid \overrightarrow{s}) \parallel Gam(\lambda \mid \alpha, \beta))$
• $err_n = D_{KL}(P(\lambda \mid \overrightarrow{s}) \parallel Gam(\lambda \mid \alpha', \beta'))$
// minimise KL-divergence
• if $err_n < err$
$err \leftarrow err_n$
$\eta \leftarrow 0$, $\beta \leftarrow \beta'$, $\alpha \leftarrow \alpha'$
• else
$\eta \leftarrow \eta + 1$

Equation 6.2 is an infinite mixture of gamma distributions. If the sensors were reliable, the mixtures would be reduced to a single gamma distribution, and the equation would simplify, by definition, to a single gamma posterior for a FOPP. Thus, it could be assumed that the sensors have minor unreliability and use a single gamma as an approximation to the posterior. It is expected that this approximation deteriorates if the sensors become more unreliable. Thus, in this approximate filter, as observations arrive, the true posterior as a sum of gamma distributions is calculated using Equation 6.2 and then fit a single gamma to it by hill climbing algorithm on the KL-divergence. Assuming a single Gamma distribution

$Gam(\lambda \mid \alpha, \beta)$ is sufficiently representative of $P_G(\lambda \mid \overrightarrow{s})$, one would expect that the Gamma distribution and $P_G(\lambda \mid \overrightarrow{s})$ behave similarly in terms of KL-divergence over λ.

To approximate $P_G(\lambda \mid \overrightarrow{s})$, an optimisation algorithm which minimises KL-divergence between a single Gamma distribution $Gam(\lambda \mid \alpha, \beta)$ and $P_G(\lambda \mid \overrightarrow{s})$ is proposed and is shown in Algorithm 2. The KL-divergence $D_{KL}(P(\lambda \mid \overrightarrow{s}) \parallel Gam(\lambda \mid \alpha, \beta))$ is chosen in this direction to specifically measure the information lost when $Gam(\lambda \mid \alpha, \beta)$ is used as an approximate of $P(\lambda \mid \overrightarrow{s})$. Since $P(\lambda \mid \overrightarrow{s})$ represents a precisely calculated distribution of λ, while $Gam(\lambda \mid \alpha, \beta)$ represents an approximation of $P(\lambda \mid \overrightarrow{s})$, in order to find a distribution $Gam(\lambda \mid \alpha, \beta)$ that is closest to $P(\lambda \mid \overrightarrow{s})$, KL-divergence must be minimised.

The algorithm operates as follows. The input of the algorithm is the posterior distribution $P_G(\lambda \mid \overrightarrow{s})$ obtained by calculating Equation 6.2, and the β parameter of the prior distribution $P_G(\lambda)$ which is assumed to follow a gamma distribution. In each step, a new β which is close to the current β is chosen, and the new parameter α of the gamma distribution is derived from the new β and the peak (the mode) of the posterior distribution $P_G(\lambda \mid \overrightarrow{s})$. The approximated gamma is then built, and the KL-divergence between the approximated gamma and $P_G(\lambda \mid \overrightarrow{s})$ is calculated. If the current divergence is smaller than the previous stored divergence, then the current α, and β of the gamma distribution are recorded. This process continues until there is no α, and β which can provide smaller divergence to $P_G(\lambda \mid \overrightarrow{s})$. Each step where the current α and β give the smallest divergence to $P_G(\lambda \mid \overrightarrow{s})$, an epoch value η increases. Whenever η goes above η_{max}, the whole process stops and the current parameters α, and β are used as the gamma approximation to $P_G(\lambda \mid \overrightarrow{s})$.

Algorithm 2 is a simple derivation of the hill climbing algorithm. The hill climbing algorithm is chosen not only due to its simplicity in the implementation but also because of its ability to find optimal solution for convex problems where this approximation belongs to. One should note that any optimisation algorithm where the objective function is to minimise KL-divergence between a single Gamma distribution $Gam(\lambda \mid \alpha, \beta)$ and $P_G(\lambda \mid \overrightarrow{s})$ can replace Algorithm 2.

This problem of approximating a mixture of gammas is entirely different from the well-

known problem of estimating the sum of a several i.i.d. variables, each of which follows a gamma distribution. The latter estimate can be approximated by the Welch-Satterthwaite approximation to a single Gamma distribution Satterthwaite (1946), Welch (1947), whereas, the former estimate has no well-known closed analytical form.

6.3.2 Histogram filter

Algorithm 3 *Histogram filter*

Input: α, β: $Gam(\lambda \mid \alpha, \beta)$
Output: (λ_i, p_i): $p_i = Gam(\lambda = \lambda_i \mid \alpha, \beta)$
Procedure:
// Initialise 0.001 (low) and 0.999 (high) percentile
// Inverse of Cummulative Distribution Function of gamma
1. Init. $low \leftarrow CDF^{-1}(\% = 0.001 \mid \alpha, \beta)$, $high \leftarrow CDF^{-1}(\% = 0.999 \mid \alpha, \beta)$
// Initialise bin, and step
2. Init. $i \leftarrow 0, \eta \leftarrow n, step \leftarrow \dfrac{(high - low)}{\eta}$
3. Repeat until $low > high$
   - $\lambda_i \leftarrow low$
   - $p_i \leftarrow Gam(\lambda = low \mid \alpha, \beta)$
   - $low \leftarrow low + step, i \leftarrow i + 1$

The Gamma filter is expected to worsen as an approximation to the posterior belief for λ when sensors become more unreliable. A histogram filter for $P(\lambda \mid \overrightarrow{x})$ may be used instead, by quantising λ. For that, the use of Gamma-Poisson conjugacy in all Equation 6.2 needs to be replaced with:

$$P_{HF}(\lambda \mid \overrightarrow{x}) = \frac{P(x_j \mid \lambda)\ P(\lambda \mid \overrightarrow{x_{-1}})}{\int_{\lambda_i} P(x_j \mid \lambda_i)\ P(\lambda_i \mid \overrightarrow{x_{-1}})\ d\lambda} \tag{6.6}$$

and redefine as Equation 6.4 as:

$$P_{HF}(x_j \mid \overrightarrow{x_{-1}}) = \int_{\lambda_i} P(x_j \mid \lambda)\ P(\lambda \mid \overrightarrow{x_{-1}})\ d\lambda \tag{6.7}$$

with $\overrightarrow{x_{-1}} = x_{j-1}, \ldots, x_1$.

Unlike the gamma filter, there is no approximation step to $P(\lambda \mid \overrightarrow{x})$ in the histogram

filter. Quantising the prior distribution in the form of gamma, which is of a continuous variable, to a series of non-overlapping bins η (or bar graph) is the approximation process itself. As the Algorithm 3 describes, the approximation process happens only if the prior distribution is in the form of gamma. As bins are used to represent the prior distribution, two equations above accommodate the loss of Gamma-Poisson conjugacy which is had in original Equation 6.2. It is expected that the higher the η value (η represents the number of bins used), the more precise the histogram filter tracks the posterior with the cost of increasing computational time.

6.3.3 Switching filter

The gamma filter is fast and can approximate the posterior well when sensors are relatively reliable. The histogram filter is slower, but can track the posterior well regardless of sensor reliability with enough computation. This leads to the possibility of a switching filter. This runs either one of the filters, using a switching mechanism in each posterior update to determine which one is used. Given the latest sampled sensor vector $\overrightarrow{s_i}$, the posterior $P(\lambda \mid \overrightarrow{s_i})$ is calculated by Equation 6.2. The gamma filter then returns an approximation to this. If the KL-divergence D_{KL} exceeds a threshold θ then the histogram filter takes over. The process repeats whenever a new sampled sensor vector $\overrightarrow{s_i}$ is given. As observations accumulate and the posterior tightens, typically, the gamma filter starts to take over and the histogram filter is switched off. This ensures that the gamma filter only replaces $P(\lambda \mid \overrightarrow{s_i})$ with a single gamma density whenever $P(\lambda \mid \overrightarrow{s_i})$ resembles a gamma distribution. Hence, the general dissimilarity between the estimated and the true posterior $P(\lambda \mid \overrightarrow{s_i})$ is minimised, while unnecessary computation is minimised.

6.4 Evaluation on Synthetic Data

To show the benefit of PoPP over FOPP in correctly estimating Poisson arrival rate λ, An evaluation to both filter and sensor model behaviour is first conducted on synthetic data.

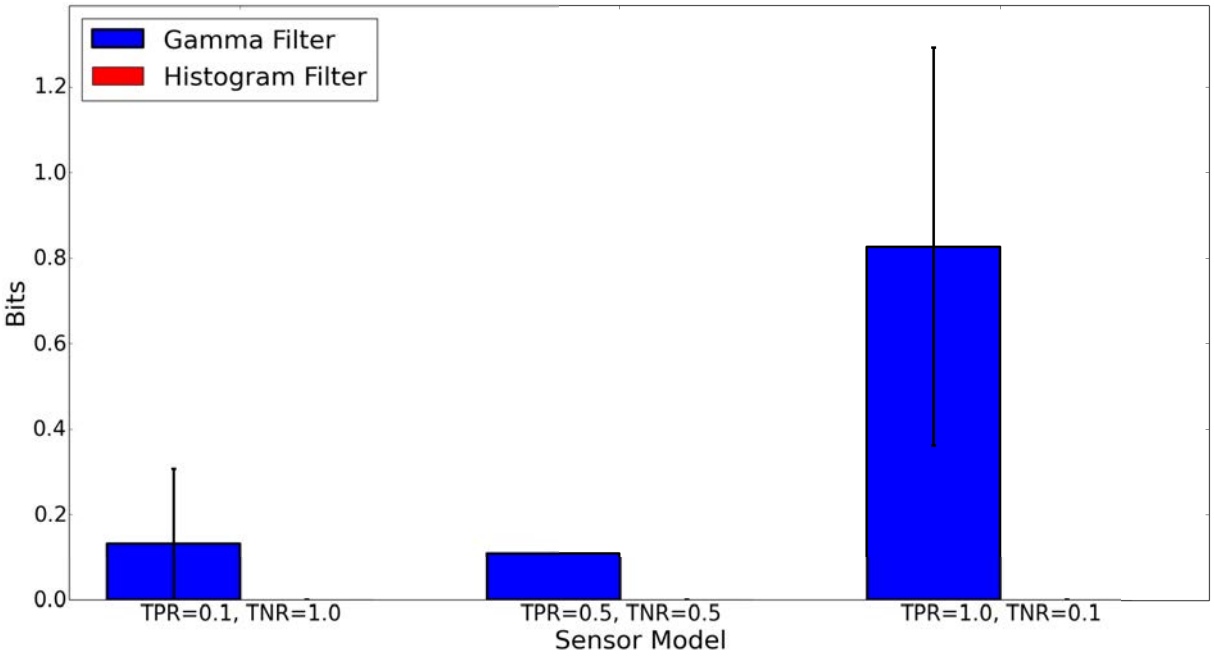

Figure 6.3: Average KL-divergence (in bits) from the gamma and histogram filters to the true posterior $P(\lambda \mid \overrightarrow{s_1}, \ldots, \overrightarrow{s_n})$ with variations on the sensor model. Standard error is shown

With synthetic data, sensor reliability can be controlled, and the true λ and the true counts x_i are known in advance for each sample. Filter performance is measured in five different ways as discussed in the following subsections.

6.4.1 Comparing filters to the true posterior

First, how different the approximations from two filters (gamma filter and histogram filter) to the true posterior in terms of KL-divergence are measured given a sequence of samples from a Poisson distribution. Since the true posterior takes an exponential amount of time to compute, this can only be achieved for very short sequences of samples from the Poisson. Seven samples were set in this evaluation due to time limitation.

The gamma filter and histogram filter are tested on three different levels of sensor unreliability:

- True positive rate: 0.1, true negative rate: 1.0;
- True positive rate: 1.0, true negative rate: 0.1;

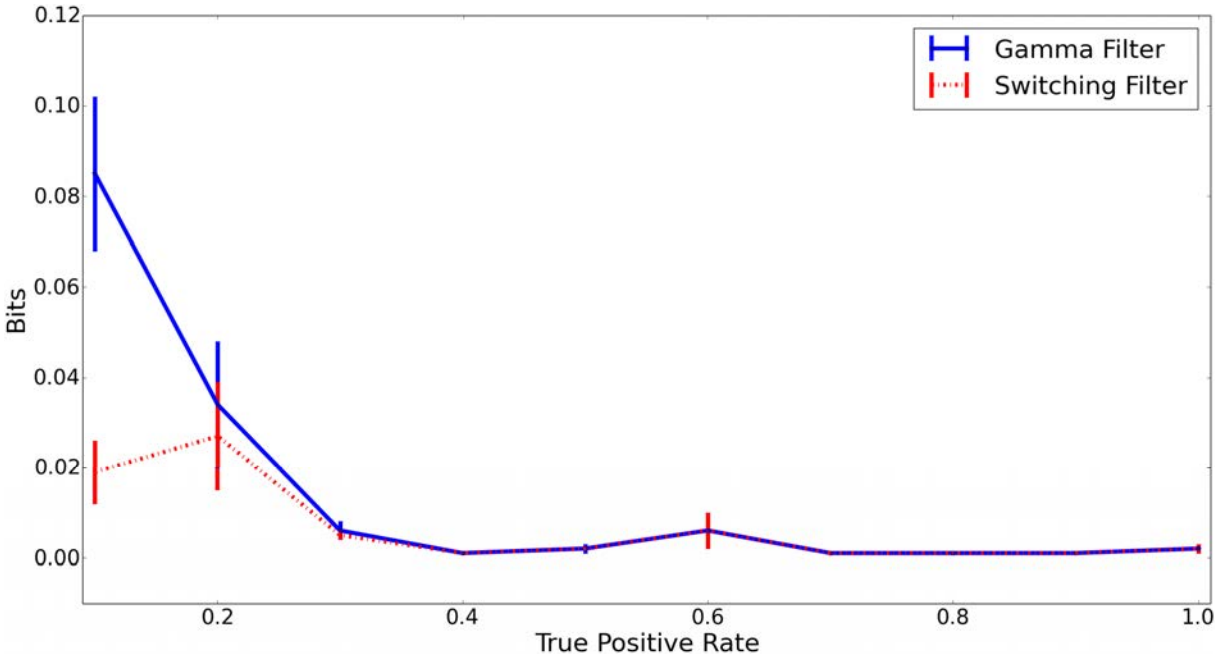

Figure 6.4: Average KL-divergence (after 144 sample counts taken from 60 trial runs) from the gamma and switching filters to the histogram filter which acts as the proxy ground truth posterior of $P(\lambda \mid \vec{s_1}, \ldots, \vec{s_n})$. The horizontal axis shows the variation in the true positive rate of the simulated sensor. Standard error is shown.

- True positive rate: 0.5, true negative rate: 0.5

to represent all extreme cases of sensor unreliability. A uniform-like prior distribution with $Gam(\lambda \mid \alpha = 1.01, \beta = 0.01)$ was used for an uninformative prior to avoid bias coming from the prior.

Figure 6.3 shows the divergence in bits. The histogram filter perfectly tracks the true posterior (zero error). The gamma filter tracks the true posterior well as long as the TNR is not extremely low. It shows that the gamma filter is a poor fit to the posterior when the sensor produces many false positives, but copes well otherwise.

6.4.2 Comparing filters on long sequences

As the previous subsection shows, the histogram filter perfectly tracks the true posterior. Thus, it can be assumed that the histogram filter will continue tracking the true posterior with very low margin of error. With that assumption, the histogram filter can be used as a *proxy ground truth posterior* or in short *proxy posterior*. This allows a comparison performance of

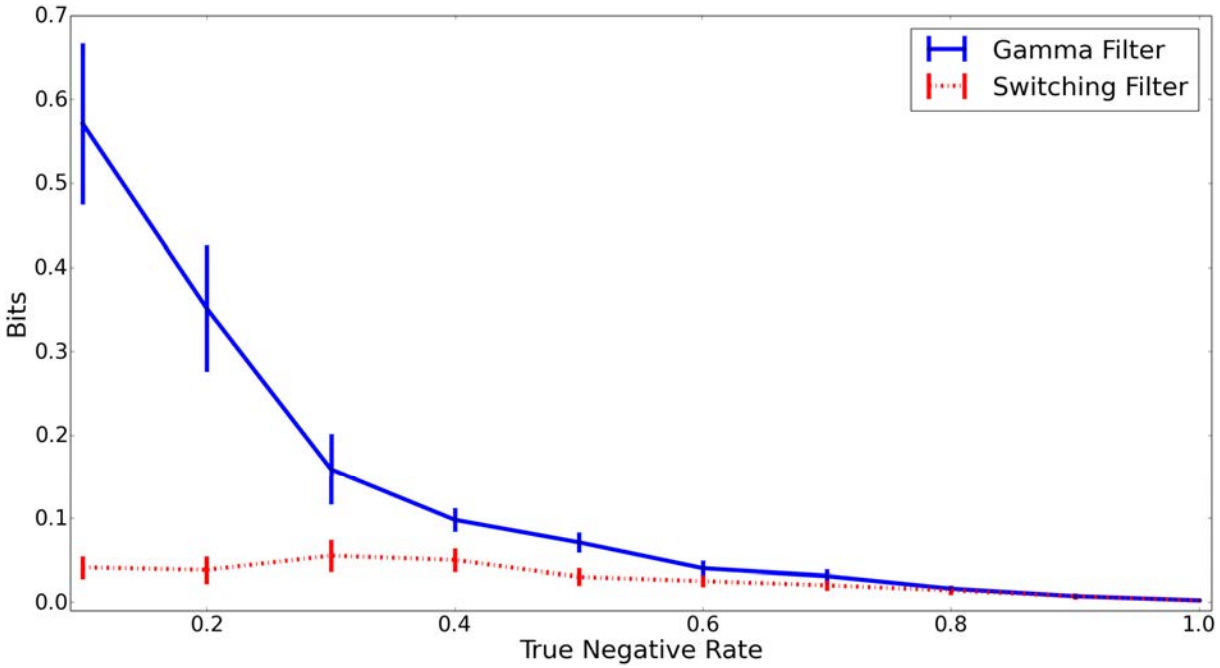

Figure 6.5: Average KL-divergence (after 144 sample counts taken from 60 trial runs) from the gamma and switching filters to the proxy ground truth posterior $P(\lambda \mid \overrightarrow{s_1}, \ldots, \overrightarrow{s_n})$. The horizontal axis shows the variation in the true negative rate of the simulated sensor. Standard error is shown.

the gamma and switching filters on much longer sample sequences.

For each trial, simulated counts $x_1, \ldots, x_{144}$ were sampled from a homogeneous Poisson process with $\lambda = 3$. These counts were then fed to a simulated sensor that counted unreliably. $P_G(\lambda \mid \overrightarrow{s_i})$ is then recursively updated using the gamma and switching filters. Repeating the process to 60 trials, the average KL-divergence from the gamma and switching filters to the proxy posterior were measured. The average for each level of sensor unreliability are shown in Figures 6.4 and 6.5.

As shown in figures above, the switching filter has a lower divergence, and the advantage over the gamma filter improves as the sensor becomes more unreliable. Please note that the gamma filter is adversely affected by a low TNR.

6.4.3 Computational efficiency

As shown in the previous subsection, the gamma filter does not always approximate the posterior as well as the histogram filter. However, it performs much faster than the histogram

Table 6.1: Comparison of the update time for three filters at each observation. η is the number of bins (histogram filter) or epoch (gamma filter) used on each observation.

Method	Average update time (std. deviation)		
	$\eta = 100$	$\eta = 1000$	$\eta = 10000$
Gamma filter	0.209s (0.031)	1.086s (0.329)	9.308s (3.624)
Switching filter	0.423 (0.455)	4.655 (3.237)	85.932s (71.387)
Histogram filter	1.982s (0.220)	16.395s (2.494)	338.374 (1.957)

filter. While the histogram filter requires a probability density function quantised as a set of points, the gamma filter has it in a closed analytical form. The computation time per sample is then examined for each of the three filters. Table 6.1 displays how the run-time for the three filters grows, as a function of the control parameters of each filter. The measurements were tested on a laptop with an i7-4702MQ Intel Core processor at 2.2 Ghz and 16 gigabytes of RAM.

In general, the gamma filter is faster than the switching filter and much faster than the histogram filter. The switching filter shows that it has high deviation on its computation time due to the switching mechanism on each update. Whenever the gamma filter fails to meet the threshold θ for the switching filter, an extra time is needed to do the histogram filter. For the remaining of this chapter, to balance accuracy against computational cost, $\eta = 100$ was set for the gamma filter, and $\eta = 1000$ was set for the histogram filter.

6.4.4 θ selection of switching filter

θ is the acceptable KL-divergence in a single filtering step in switching filter from the true posterior (or the histogram filter as a proxy) to the approximation. The selected θ influences how low the KL-divergence is and how fast the computation time per sample is. Previous sections, θ was chosen to be 0.05 for the switching filter. In this subsection, a systematic evaluation of the effect of θ on computation time and KL-divergence over long runs is performed, and the decision to choose 0.05 to be the θ is explained. As shown in Figures 6.4 and 6.5, it is low TNR values that affect the gamma approximation most. Hence, the TNR of a single sensor is varied for this experiment.

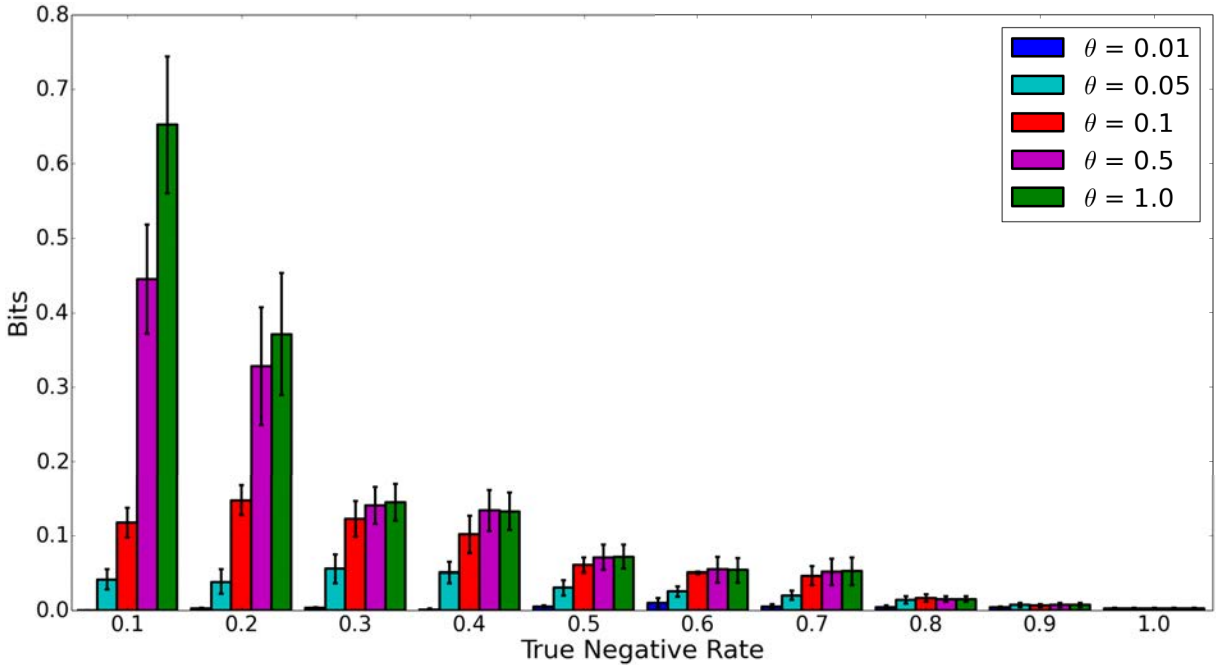

Figure 6.6: Average KL-divergence from the switching filter to the proxy ground truth posterior $P(\lambda \mid \overrightarrow{s_1}, \ldots, \overrightarrow{s_n})$. The horizontal axis shows the variation in the true negative rate of the simulated sensor.

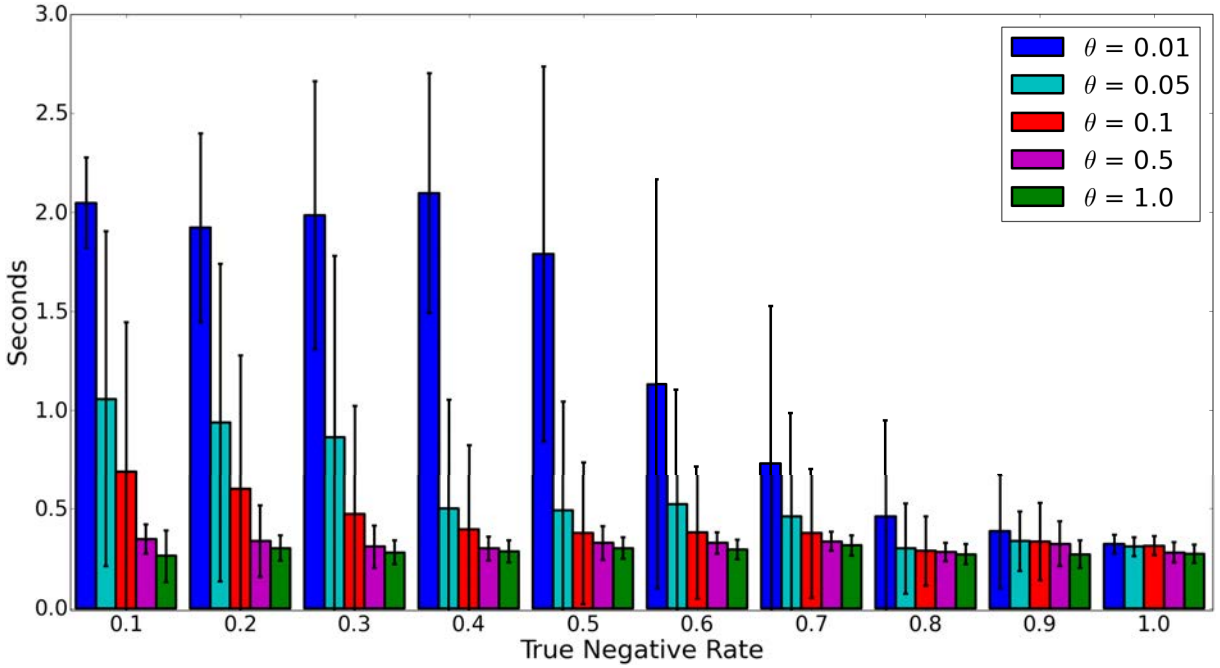

Figure 6.7: The updating time per sample for the switching filter (with $\eta = 100$ both for the histogram filter and the gamma filter).

Each trial, simulated counts $x_1, \ldots, x_{144}$ by random sampling (Poisson with $\lambda = 3$) were generated and fed to the simulated sensor. $P(\lambda \mid \overrightarrow{s_i})$ was updated using only the switching

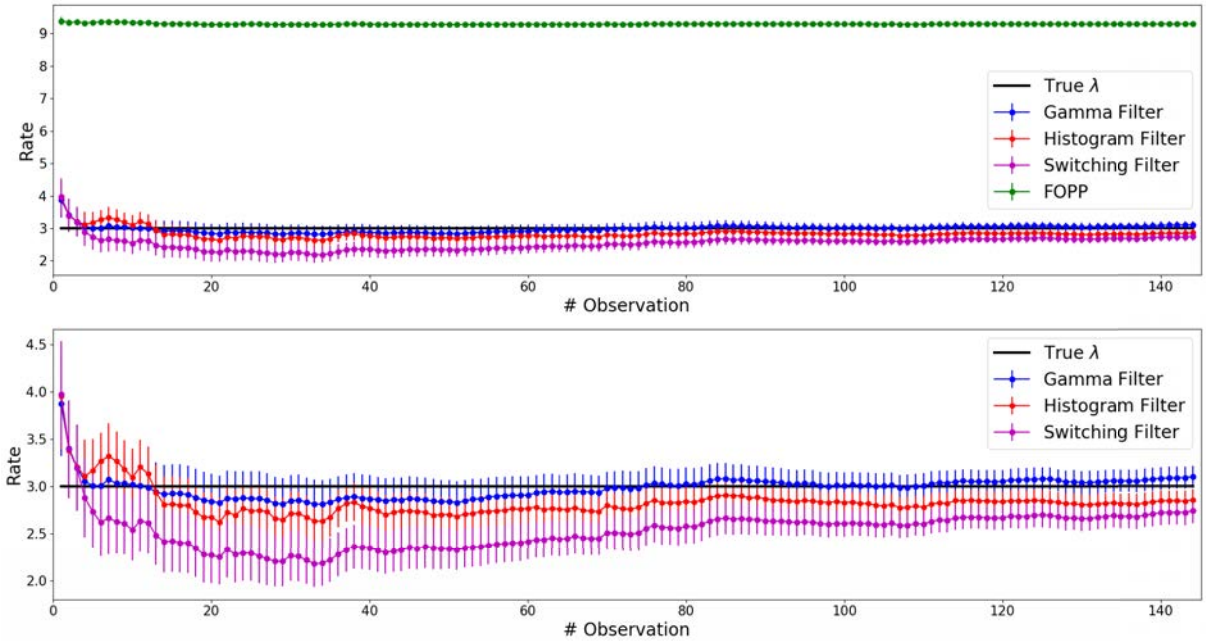

Figure 6.8: The evolution of λ posterior for each filter with sensor tpr = 1.0, and tnr = 0.1. The posterior is presented as MAP estimates. Top figure includes the FOPP filter, whereas the bottom figure excludes it. Each data point is an average of 30 trials. Standard errors are shown.

filter with variation on θ. 30 trials were run, and these show the average divergence and calculation time for each level of TNR reliability in Figures 6.6 and 6.7 respectively.

Figures 6.6 and 6.7 show how five different values for θ (0.01, 0.05, 0.1, 0.5 and 1.0) change the divergence and computation as the TNR changes. For, $\theta = 0.05$ the total bit divergence after a simulated run of 144 steps with a (True Negative Rate) TNR of 0.1 and a single sensor is around 0.04 bits. This can be further reduced by setting $\theta = 0.01$ to 0 bits, but in that case the switching filter only ever runs the histogram filter.

6.4.5 Accuracy of posterior estimates of λ

As the quality of the posterior for each filter has been described and evaluated, it is now the time to examine the performance accuracy of estimating λ value relative to an incorrect assumption of sensor reliability. In other words, a comparison of each filter of POPP in estimating λ relative to filtering according to a FOPP model was performed. This experiment is conducted for both single and multiple sensors at each sample.

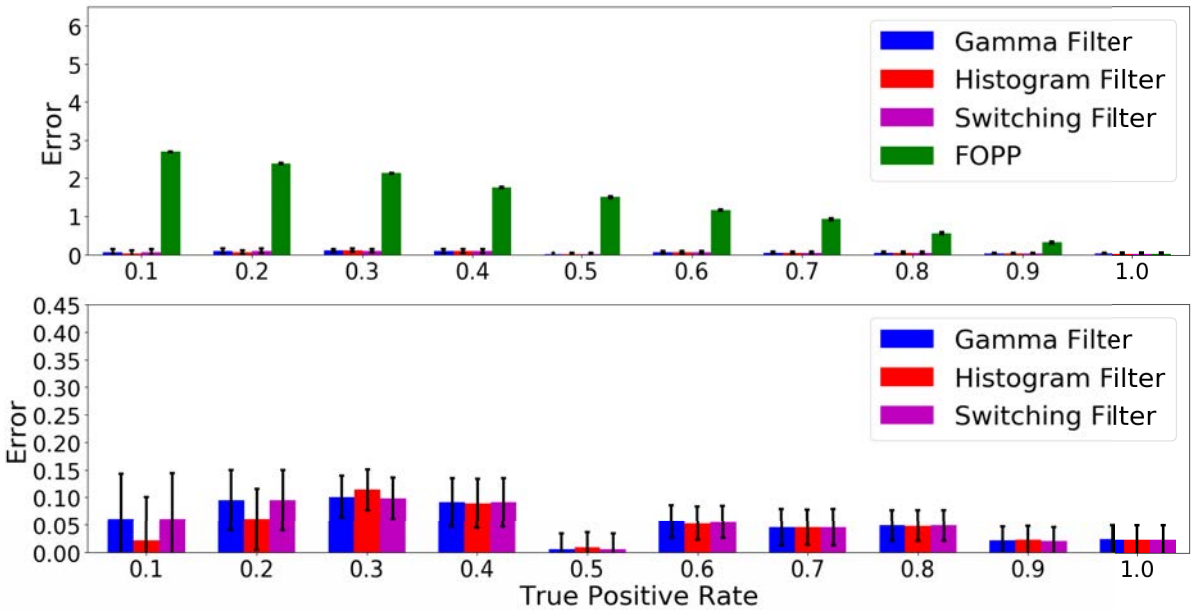

Figure 6.9: The RMSE of posterior estimates of λ for different filters. In each graph the unreliability of a single sensor is varied. Each trial consisted of a stream of $\overrightarrow{s_1} \ldots \overrightarrow{s_{144}}$ samples. Estimation accuracy for λ while varying the TPR. Accuracies include the FOPP filter in the top panel for a comparison, and exclude the FOPP filter in the bottom panel. Each data point is an average of 30 trials. Standard errors are shown.

To make this comparison, the corrected estimate of the arrival rate λ of a Poisson process is evaluated by applying Equation 6.2, using all three filters for the POPP model. This is compared against the uncorrected estimate using standard inference according to the FOPP model. In each experiment, 144 simulated counts $x_1, \ldots, x_n$ were sampled from a Poisson process $P(x \mid \lambda')$ together with their corresponding sensor readings. $P(\lambda \mid \overrightarrow{s_i})$ is then recursively updated using three filters for the POPP model, and Equation 6.1 for the FOPP model. Figure 6.8 shows an example of the evolution of the posterior of λ for each filter on each sample update.

Sensor specificity (true negative rate) is varied while fixing sensor sensitivity (true positive rate) and vice versa. The performance of both POPP filters and a standard FOPP filter were assessed by comparing the RMSE of MAP hypothesis (mode) of each posterior over λ to the true λ'. The results for a filter with a single sensor input are shown in Figures 6.9 and 6.10. This shows that all POPP filters generate much better estimates of λ than the FOPP filter. The FOPP filter's performance decreases linearly as sensor reliability declines.

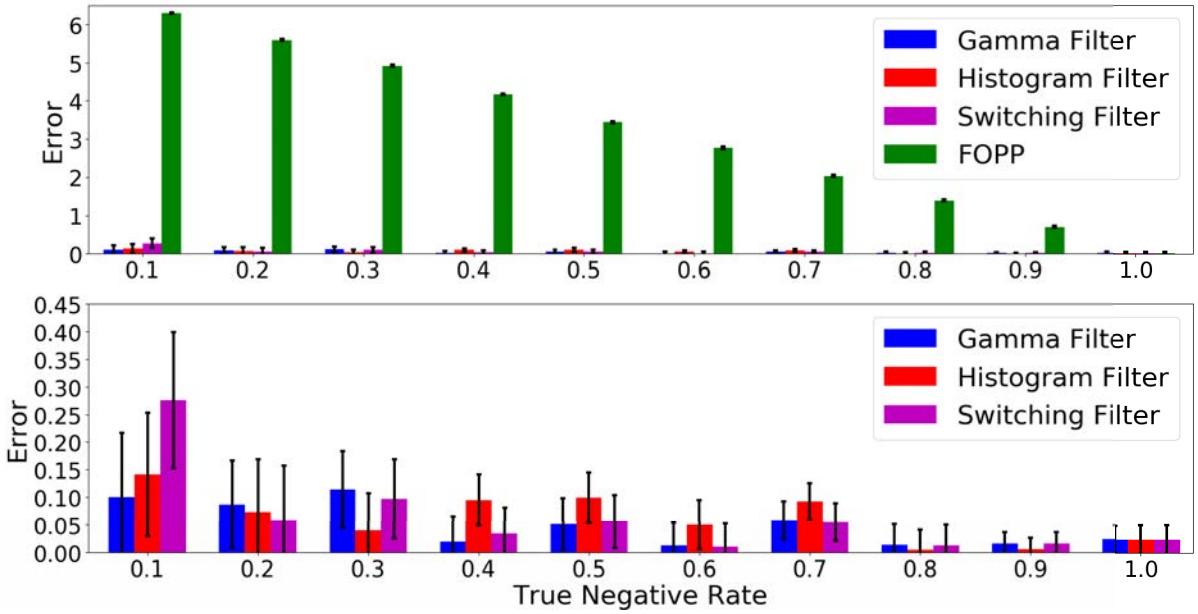

Figure 6.10: The RMSE of posterior estimates of λ for different filters. In each graph the unreliability of a single sensor is varied. Each trial consisted of a stream of $\overrightarrow{s_1} \ldots \overrightarrow{s_{144}}$ samples. Estimation accuracy for λ while varying the TNR. Accuracies include the FOPP filter in the top panel for a comparison, and exclude the FOPP filter in the bottom panel. Each data point is an average of 30 trials. Standard errors are shown.

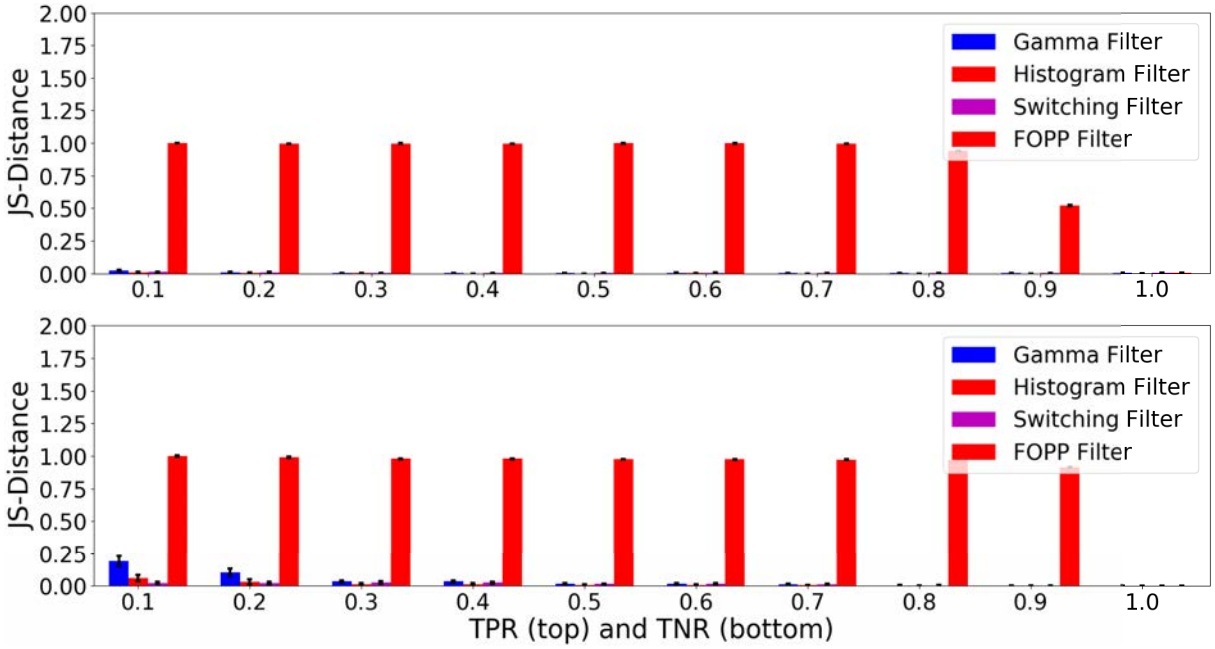

Figure 6.11: The Jensen-Shannon distance of λ for the FOPP, gamma, and switching filters with variation on the number of sensors. In each graph the unreliability of a single sensor is varied. Estimation accuracy for λ while varying the TPR on the top and estimation accuracy for λ while varying the TNR on the bottom. Each data point is an average of 30 trials. Standard errors are shown.

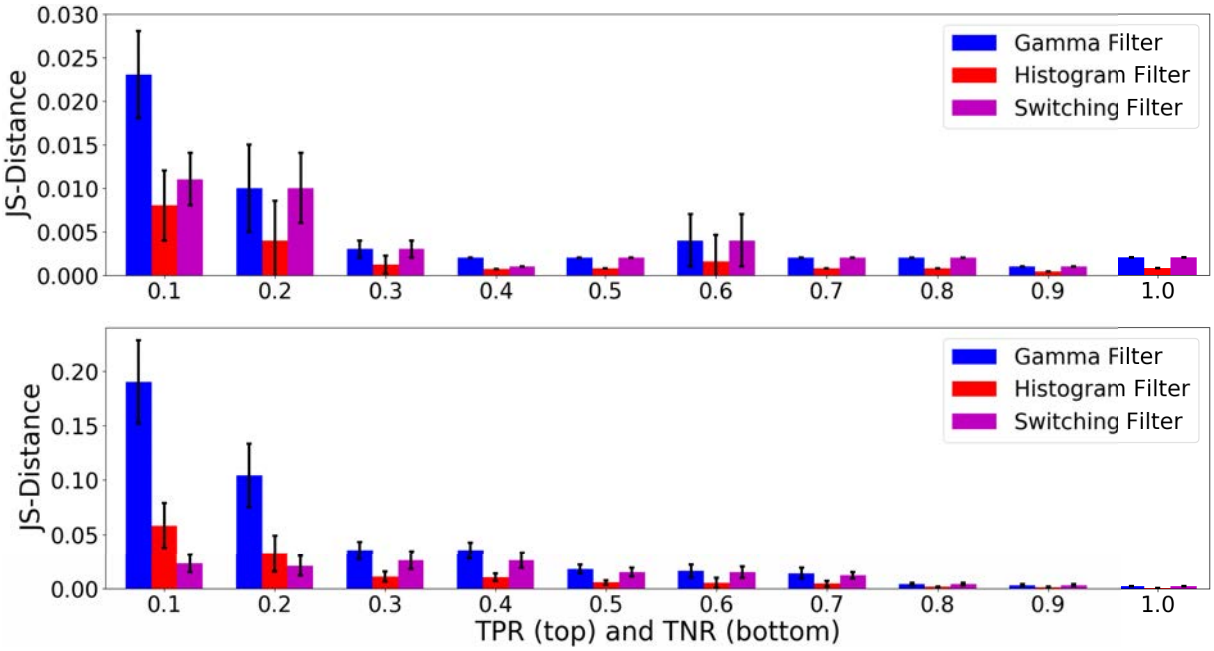

Figure 6.12: The Jensen-Shannon distance of λ for gamma, and switching filters with variation on the number of sensors. In each graph the unreliability of a single sensor is varied. Estimation accuracy for λ while varying the TPR on the top and estimation accuracy for λ while varying the TNR on the bottom. Each data point is an average of 30 trials. Standard errors are shown.

The performance of both POPP filters and the FOPP filter were also assessed by comparing the Jensen-Shannon distance. Jensen-Shannon distance is a unit of measurement used in Jensen-Shannon divergence. The Jensen-Shannon divergence is a method of measuring the similarity between two probability distributions. Unlike KL-divergence, Jensen-Shannon divergence is a symmetrized divergence where $D_{JS}(P \parallel Q)$ is equal to $D_{JS}(Q \parallel P)$. The results for the POPP filters and the FOPP filter on a single sensor are shown in Figures 6.11 and 6.12. These results support the result shown in Figures 6.9 and 6.10 that POPP filters indeed generate much better estimates of λ than the FOPP filter.

The performance of POPP filters and a FOPP filter by the number of unreliable sensors was also evaluated. The sensor models were varied in such a way that each pair of sensors have their sum of TPR and TNR equal to one (e.g. sensor A TPR=0.1, TNR=0.2, sensor B TPR=0.9, TNR=0.8). As it involves more than one sensor, the standard Poisson model displayed in the figure is the result of averaging the mode of all estimated λ from each sensor. The result is shown by Figures 6.13 and 6.14. The figure shows that (given this balancing) the number of

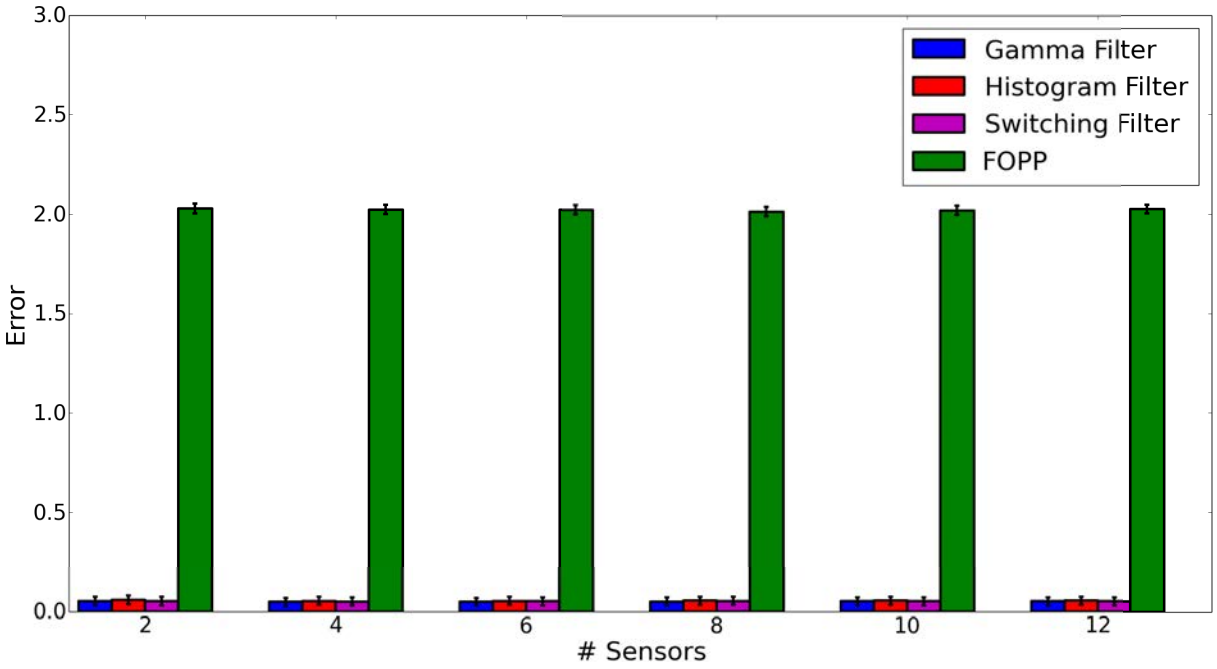

Figure 6.13: The RMSE of posterior MAP estimates of λ for the FOPP, gamma, histogram and switching filters with variation on the number of sensors. Each trial consisted of a stream of $\overrightarrow{s_1} \ldots \overrightarrow{s_{144}}$ samples. Each data point is an average of 30 trials. Standard errors are shown.

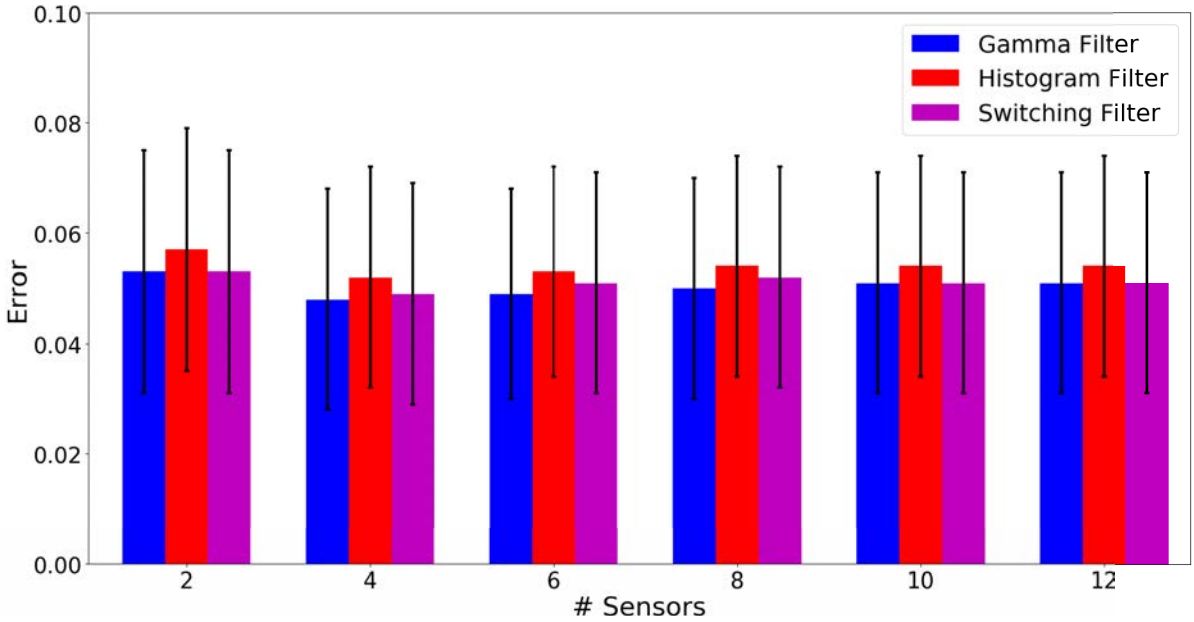

Figure 6.14: The RMSE of MAP estimates of λ for gamma, histogram and switching filters with variation on the number of sensors. Each trial consisted of a stream of $\overrightarrow{s_1} \ldots \overrightarrow{s_{144}}$ samples. Each data point is an average of 30 trials. Standard errors are shown.

sensors has no effect on the accuracy of the estimates of λ for POPP model filters. Whereas the FOPP model filter is far from the true λ'. This conclusion is supported by the Jensen-Shannon

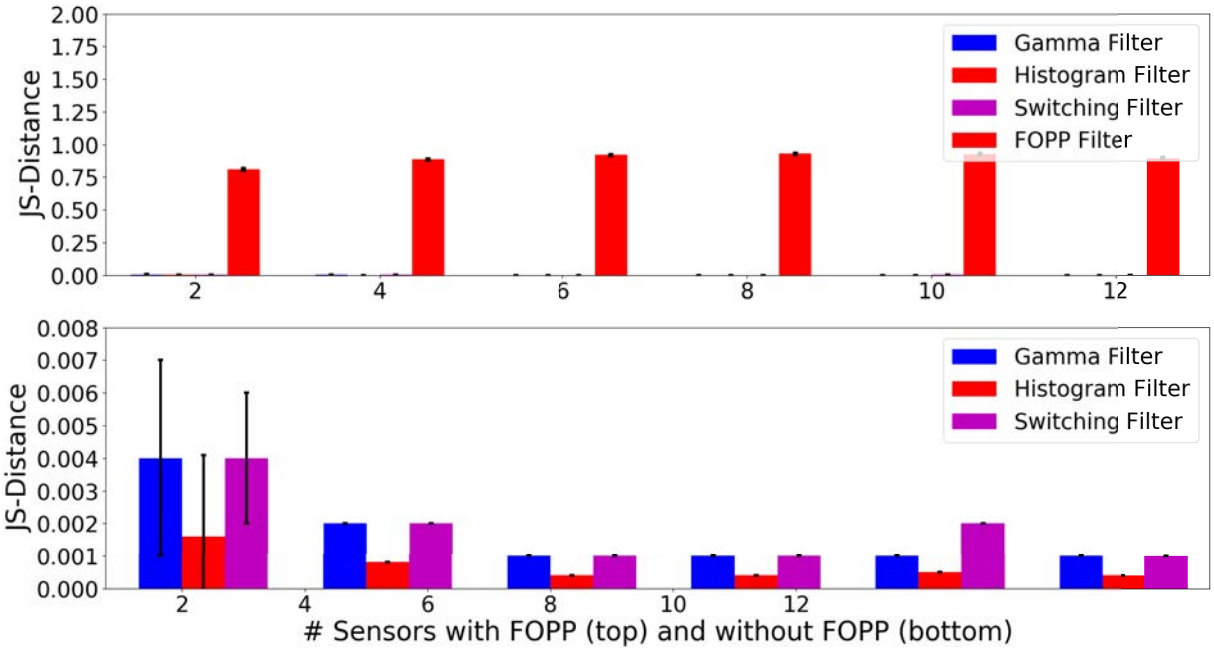

Figure 6.15: The Jensen-Shannon distance of λ for the FOPP, gamma, and switching filters with variation on the number of sensors. The top figure shows a comparison including the FOPP filter, whereas the bottom figure does not. Each data point is an average of 30 trials. Standard errors are shown.

distance of the FOPP filter which is, in fact, further to the proxy posterior of $P_G(\lambda \mid (\overrightarrow{s_1}, \ldots, \overrightarrow{s_{144}})$ than both the gamma filter and the switching filter. This is shown in Figure 6.15.

6.5 Evaluation on A Real World Dataset

Having thoroughly evaluated the performance of POPP filters on simulated data, the investigation now moves to evaluate the performance of the POPP model and its associated filters on a large, real-world data set. Performance comparison against the FOPP model is still included in this evaluation.

This dataset is a collection of counts over time from three different automated person detectors. These use laser (leg detector - LD), depth camera (upper body detector - UBD), and RGB information (change detector - CD). The dataset was gathered from an office building where a mobile robot counts and observes the number of people passing by as it patrols around the area (see Figure 6.16). Each detector/sensor returns a sensed count of the number

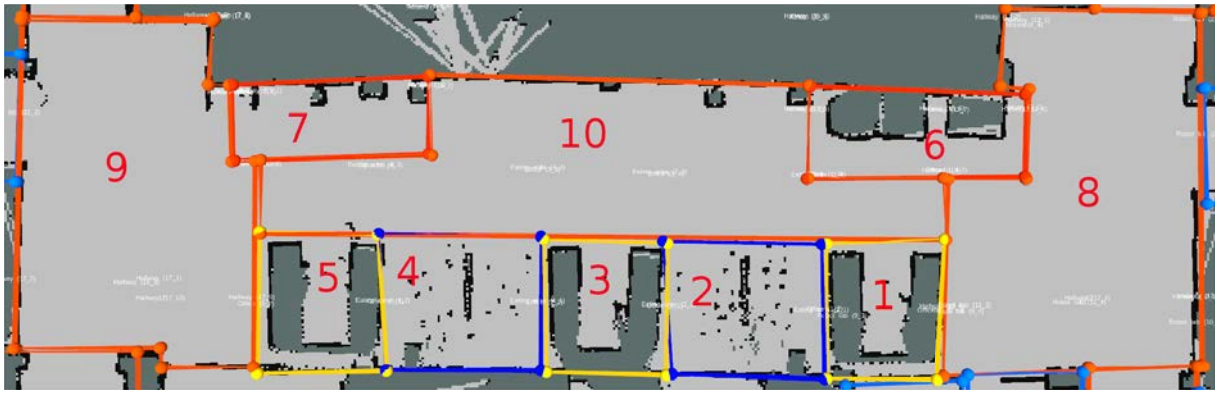

Figure 6.16: The office building in which the robot gathered data. Areas are bounded by imaginary lines.

of people it detected in each 10 minute interval during the day. As expected, these detectors are unreliable. Some examples of (true and false) positive and (true and false) negative detections made by the detectors can be seen in Figure 1.2.

The data set was gathered during a deployment of the mobile robot at the lower ground of the computer science building in University of Birmingham. As a mobile robot can not fully sense its environment, it can only perceive partial data at a particular time and place. Moreover, the robot's patrol policy also affects where and when detections are perceived. Consequently, the detections are temporally and spatially scattered and they are not uniformly distributed across space. The detections are then organised according to time/date and the spatial region where each detection was made.

During the deployment of the mobile robot, ground truth data were also gathered to create sensor unreliability models, one for each detector in each area. An example of such a sensor model can be seen in Table 6.2. The sensor models were built from the data collected from 10am-8pm each week day to obtain similar sensor model behaviour as some of the detectors suffer from the lightning condition during night patrol. The data used for this experiment was limited to weekdays to have similar temporal detection rates.

On this evaluation, periodic Poisson processes are included as a showcase comparison whether the non-homogeneous Poisson processes, in which the periodic Poisson processes are categorised into, are benefited by the POPP filters. This section starts with a comparison on homogeneous Poisson processes on a 21-day deployment dataset. The dataset was then

Table 6.2: Averaged sensor model across all areas trained from 15 days of data.

Sensor	True Positive	True Negative
Leg Detector	0.315	0.894
Upper Body Detector	0.266	0.853
Change Detector	0.611	0.820

extended to a 69-day deployment dataset to accomodate the comparison of periodic Poisson processes.

6.5.1 The POPP models on Homogeneous Poisson Processes

The dataset was gathered during a 21-day deployment of the mobile robot. As described previously, the robot was limited to observe during weekdays, 15 days of observations were obtained. The interval time for the Poisson was specified to 10 minutes, and the true counts and the detections made by each detector were recorded. The parameter λ of the Poisson distribution was estimated for each region by running a FOPP filter on the true counts. The uncorrected estimate λ according to the FOPP model was estimated only from the change detector count data since the change detector is the most reliable detector among three detectors available in the robot as shown in Table 6.2.

All POPP filters rely on sensor models which must be calculated from sensor counts and true counts. To separate the training and testing data five fold cross-validation was performed with the unit being whole days, i.e., 3-days of data was used as a training set for a sensor model and then the remaining 12-days of data was used as a test set on which to test the inferences made by each filter from the sensor counts.

Using the 12 days of test data, the filter from each model (the switching filter for the POPP model and the standard Bayes' filter for the FOPP model) built a prediction of λ based on the MAP estimate of each model. The resulting prediction from each estimator was then compared to the λ using the root mean squared error (RMSE). Moreover, the resulted distribution of each model was also compared to the distribution of the true λ using the Jensen-Shannon divergence. Using these metrics, the performance of the POPP model was

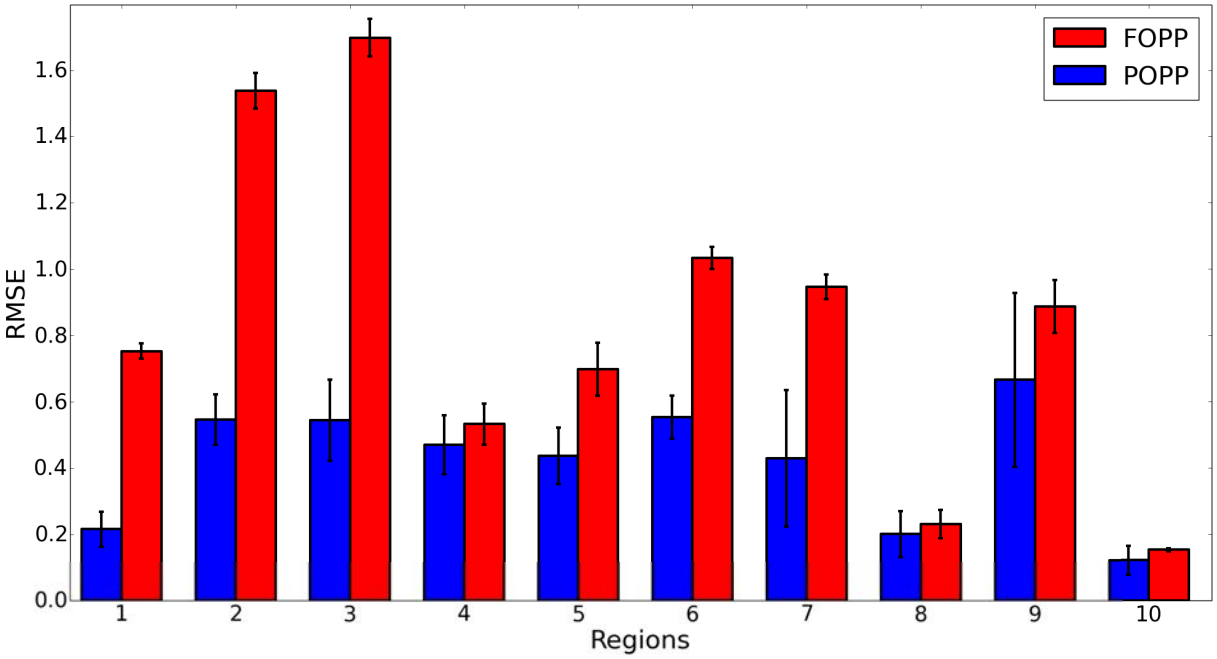

Figure 6.17: The RMSE of the FOPP and POPP estimators of λ as it varies across areas (regions) of the environment. Standard error is shown.

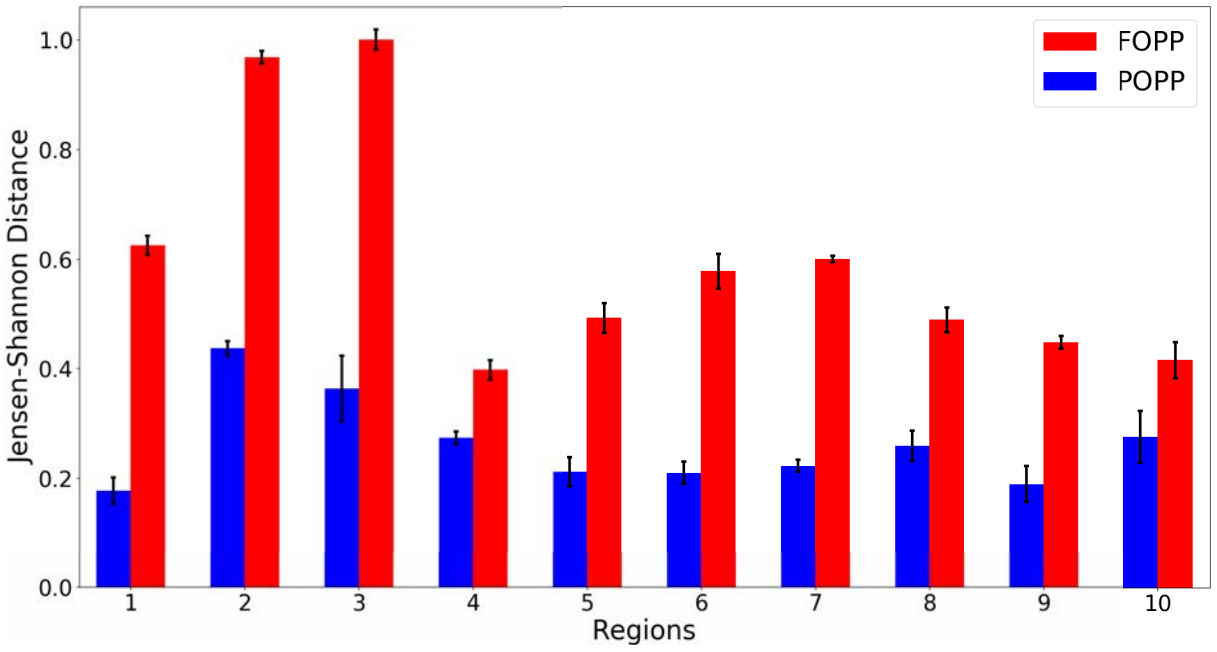

Figure 6.18: The Jensen-Shannon distance of the FOPP and POPP model distributions of λ as it varies across areas (regions) of the environment. Standard error is shown.

compared to the standard FOPP model. The comparison was done for all areas within the patrol space.

The results are shown in Figure 6.17 and 6.18. As can be seen, the standard POPP model

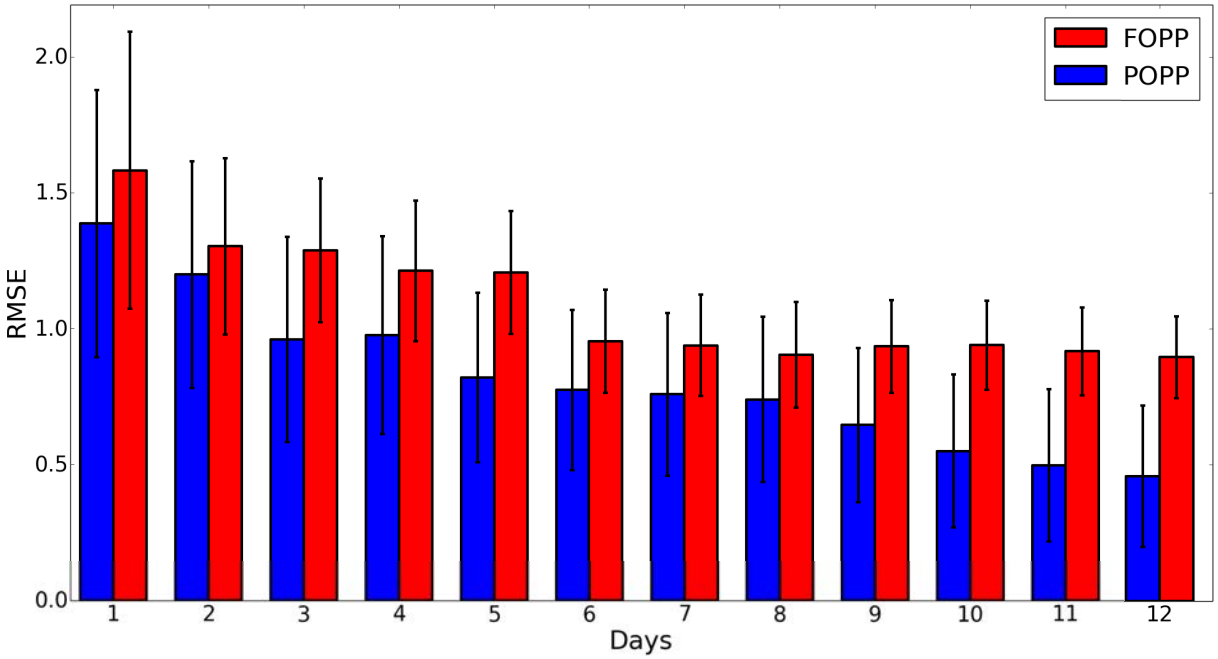

Figure 6.19: The RMSE evolution from day 1 to day 12, averaged across all regions. Standard error is shown.

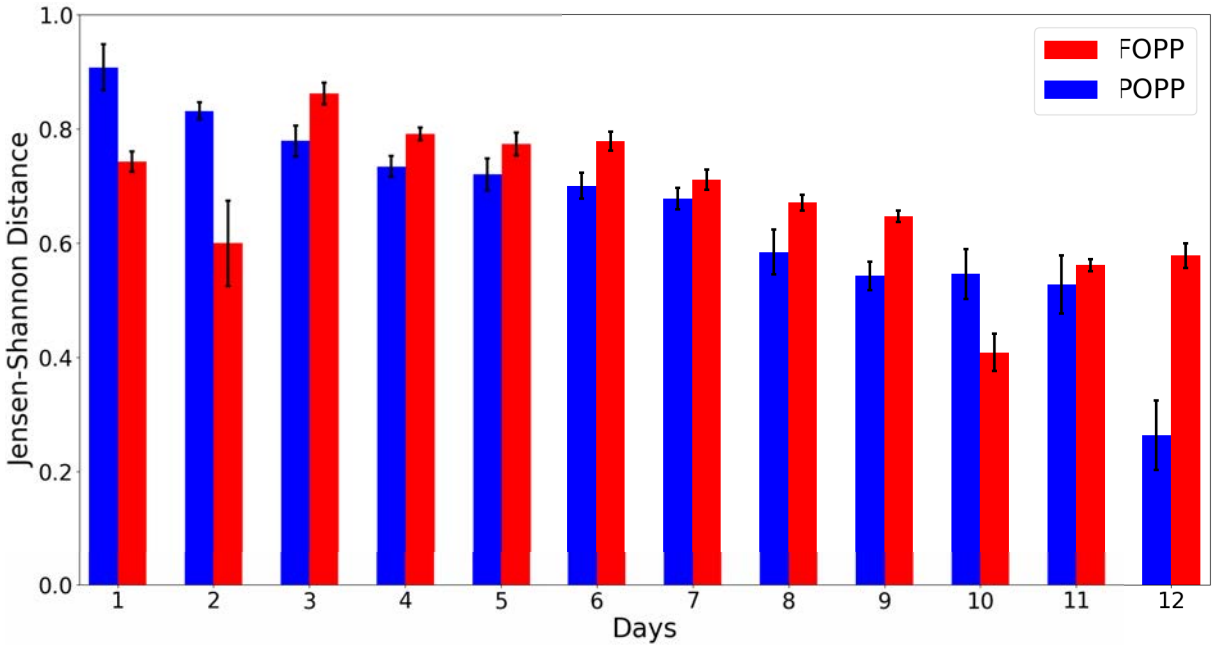

Figure 6.20: The Jensen-Shannon distance evolution of the FOPP and POPP model distributions of λ from day 1 to day 12, averaged across all regions. Standard error is shown.

with the switching filter consistently produces better MAP estimates than FOPP. Moreover, the POPP model also produces more similar distribution to the true λ than the FOPP. As mentioned earlier that the distribution of detections is not uniform across areas, some areas

Table 6.3: Averaged sensor model across all areas trained from 48 days of data.

Sensor	True Positive	True Negative
Leg Detector	0.387	0.951
Upper Body Detector	0.356	0.882
Change Detector	0.731	0.900

(such as area 4) have limited data such that the underlying sensor model for the POPP is relatively unreliable. This results in a smaller gain in both RMSE and Jensen-Shannon divergence over the FOPP estimator for those areas. The opposite was true for some areas such as 1, 2, and 3, where there are plenty of data to construct the sensor models, resulting a much better gain in both RMSE and Jensen-Shannon divergence over the FOPP estimator.

With the correction that the POPP model has, the POPP model is expected to gradually draw closer to the true λ. Here, how the POPP and FOPP estimators evolved with time is shown, again in terms of both RMSE and Jensen-Shannon distance. Figures 6.19 and 6.20 show that as time passes the performance of the POPP estimator steadily becomes better. This is shown by the growing gap between the two estimators.

6.5.2 The POPP models on Periodic Poisson Processes

For this evaluation, the dataset was extended from a 21-day deployment to a 69-day deployment of the mobile robot. Moreover, the underlying process on each region was then assumed to be a periodic Poisson process in which the imposed single periodicity is a one-day cycle, i.e. $\lambda(t_i, t_j) = \lambda(t_{i+\Delta}, t_{j+\Delta})$ with $\Delta = 24 * 60$(minutes). As the robot was limited to observe during weekdays, 48 days of observations were obtained and the sensor models were built from these observations. An example of a fixed sensor model for the POPP filter trained from 48 days of data can be seen in Table 6.3.

The true $\lambda(t_i, t_j)$ of the periodic Poisson processes for each region is estimated by running a FOPP filter on the true counts. The uncorrected estimate $\lambda(t_i, t_j)$ according to the FOPP model was estimated only from the change detector count data since the change detector is the most reliable detector among three detectors available in the robot as shown

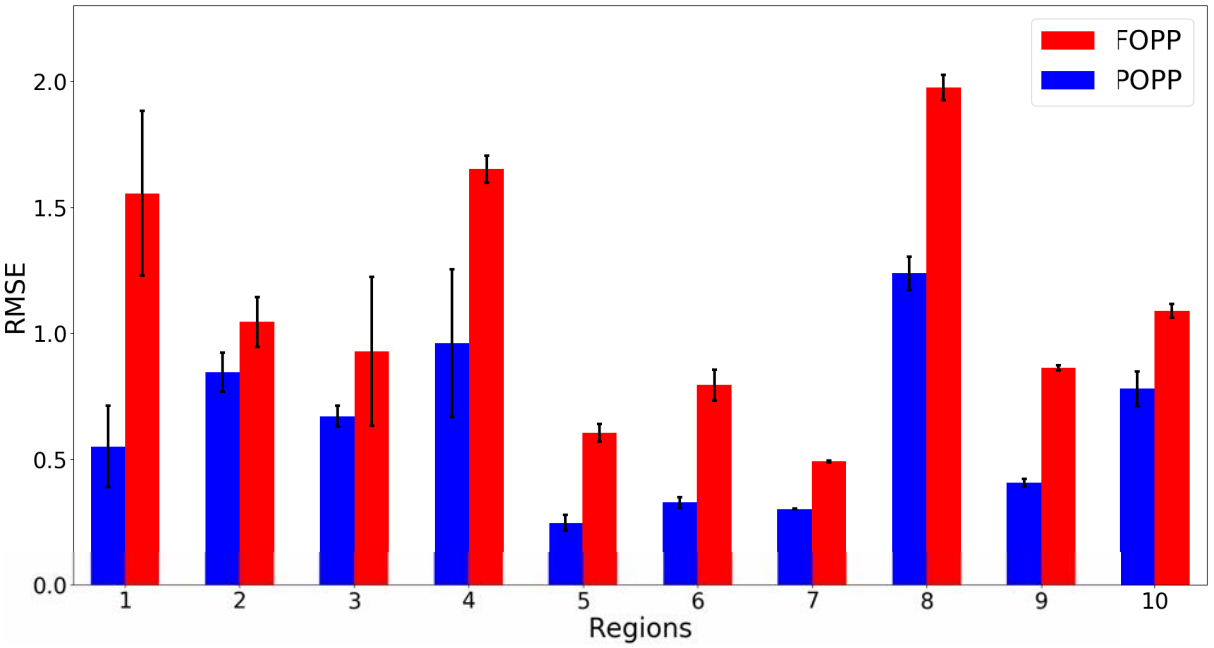

Figure 6.21: The RMSE of the FOPP and POPP estimators of $\lambda(t_i, t_j)$ as it varies across areas (regions) of the environment. Standard error is shown.

in Table 6.3.

Four fold cross-validation to separate the training and testing data were performed, i.e., 12-days of data were used as a training set for a sensor model, and the remaining 36-days of data were used as a test set to build a prediction of the $\lambda(t_i, t_j)$. The POPP model used the switching filter as its filter to do the inferences from the sensor counts.

Both point estimate predictions based on the MAP estimate and the estimated distributions of both the POPP model and the FOPP model were then compared to the true $\lambda(t_i, t_j)$ and its distribution using RMSE (see Figure 6.21) and Jensen-Shannon distance (see Figure 6.22). The comparison was done for all areas within the patrol space.

Figure 6.21 and 6.22 show that the POPP filter succeeded in gaining better predictions of $\lambda(t_i, t_j)$ over the FOPP filter both in distance metric (RMSE) and in Jensen-Shannon distance in many regions. Some areas such as regions 4, 5 and 7 have limited data such that the underlying sensor models for the POPP is relatively unreliable. This results in a smaller gain in both RMSE and Jensen-Shannon distance over the FOPP filter. On the other side, region 1 and 2 have plenty of data to construct the sensor model resulting a much better gain in both RMSE and Jensen-Shannon distance over the FOPP filter.

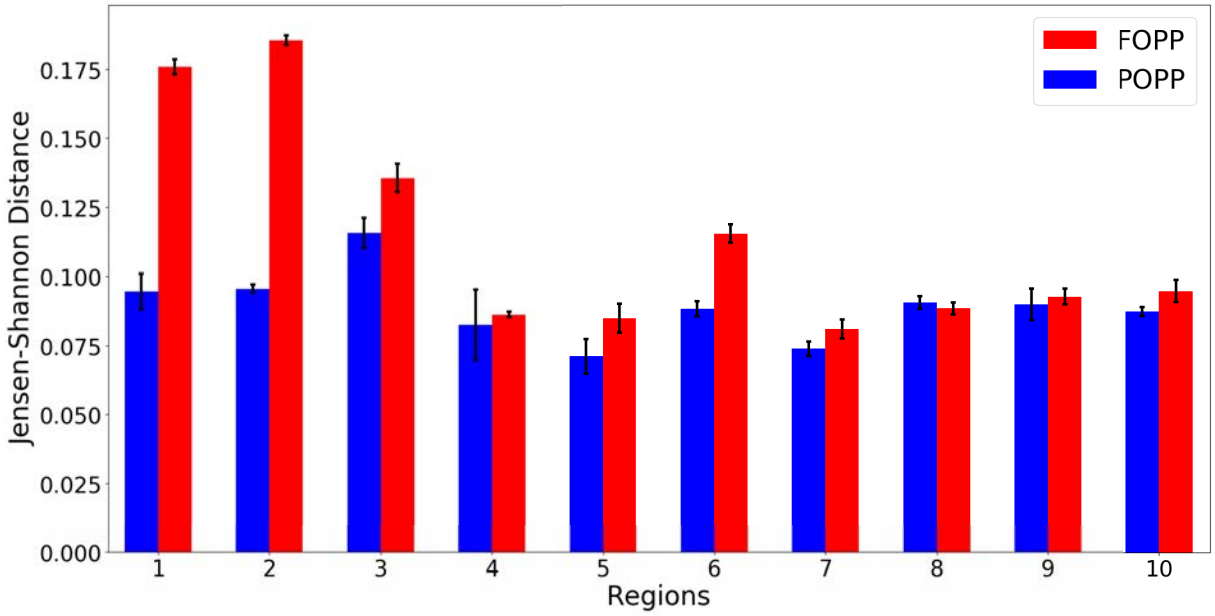

Figure 6.22: The Jensen-Shannon distance of the FOPP and POPP model distributions of $\lambda(t_i, t_j)$ as it varies across areas (regions) of the environment. Standard error is shown.

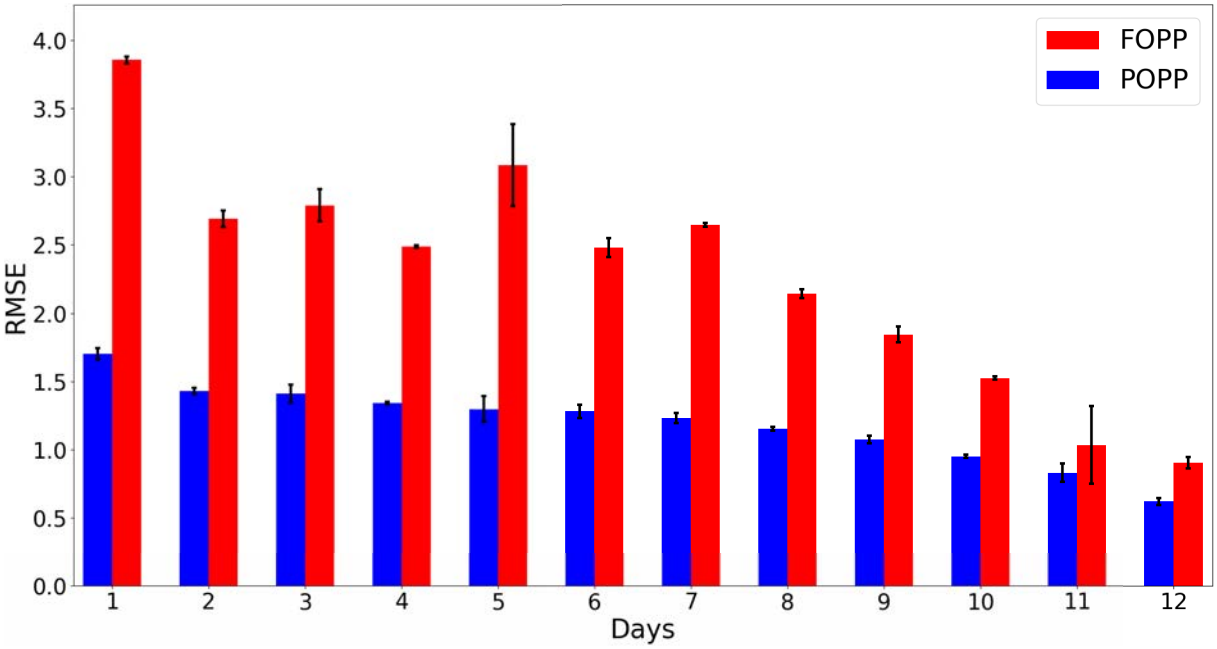

Figure 6.23: The RMSE evolution of periodic Poisson processes with POPP and FOPP filters from day 3 to day 36, averaged across all regions. Standard error is shown.

Figures 6.23 and 6.24 show the evolution of each filter overtime in terms of RMSE and Jensen-Shannon distance across regions. The POPP filter gradually increased its accuracy in estimating $\lambda(t_i, t_j)$. Although the FOPP filter also showed a gradual increase in its accuracy, the POPP filter was always more accurate than the FOPP filter.

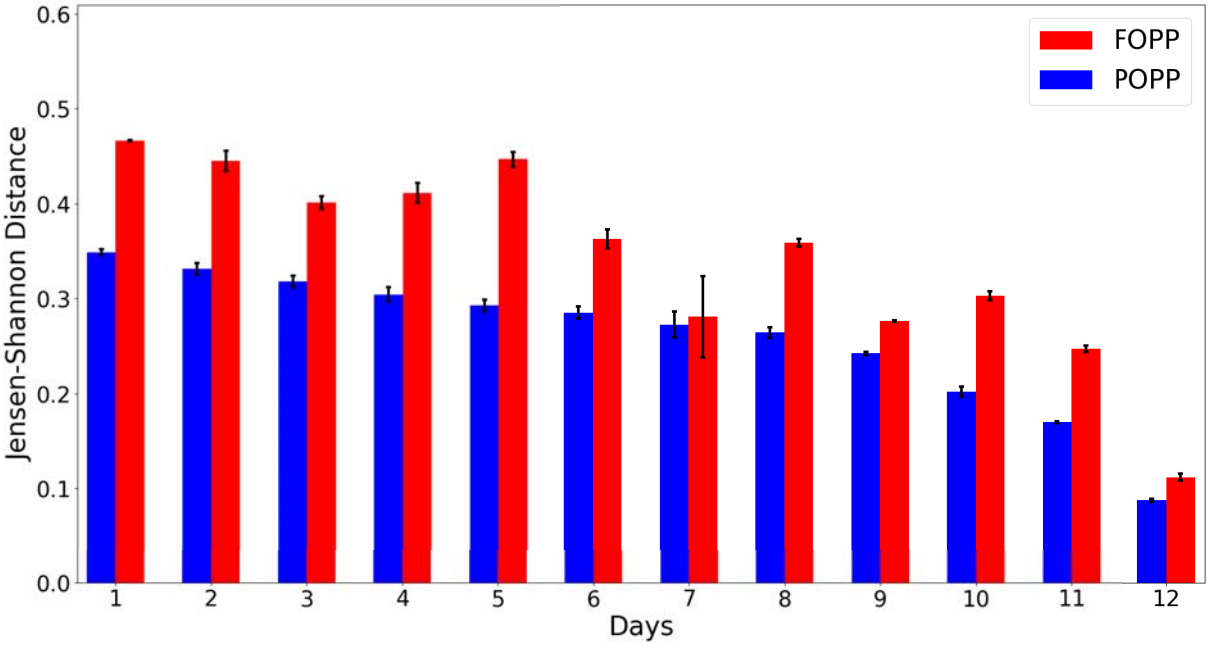

Figure 6.24: The Jensen-Shannon distance evolution of periodic Poisson processes with POPP and FOPP filters from day 3 to day 36, averaged across all regions. Standard error is shown.

6.6 Discussion

In this chapter, a model, the partially observable Poisson process (POPP), was formulated which takes into account sensor unreliability to correctly estimate the parameter of a Poisson distribution. The primary motivation is driven by the fact that there are situations where counters are extremely unreliable and the inference made from them is significantly incorrect. The problem of how to efficiently correct miscounts made by either single or multiple unreliable counting devices observing a single Poisson process is addressed.

Due to no conjugate density and the exponential growth of the number of elements in the posterior with every sample, three tractable approximations are proposed under reasonable independence assumptions. One (the Gamma filter) is fast but very dependent on the reliability of the counters. The second (the histogram filter) is slower, but it perfectly tracks the true posterior. And the third is a combination of a gamma filter and a histogram filter which avoids drift while remaining faster on average.

Experiments both on simulated data and real-world dataset (on the homogeneous Poisson process) showed that the POPP model and its associated filters outperformed a

standard Bayes' filter for the fully observable Poisson process (FOPP) model in estimating λ of a Poisson process. This chapter shows the posterior of the POPP model densely envelops the true λ, whereas the FOPP model can not achieve this without a perfect sensor model. These results, however, become less apparent on the case of periodic (non-homogeneous) Poisson processes on real-world dataset where the POPP model, in some cases/regions, perform worse than the standard FOPP model.

The POPP provides a way to deal with counts coming from unreliable counting devices. However, the POPP explained in this chapter relies on two things. First, the model here assumes that the degree of the unreliability of a sensor, regarding its TPR and TNR, is perfectly known. Second, it assumes that counter failures are conditionally independent of one to another given the true count.

CHAPTER 7

POPP: INDEPENDENT SENSORS WITH UNRELIABLE SENSOR MODELS

The previous chapter introduced the POPP model which efficiently corrects miscounts made by multiple unreliable counting devices observing a single Poisson process. The model relies on two assumptions:

- the degree of the unreliability of a sensor is precisely known, and
- counter failures are conditionally independent from one to another given the true count.

This chapter focuses on addressing the first assumption by introducing sensor unreliability as latent variables and updating Equation 6.5. Hence, the contribution here is introducing Bayesian estimators for sensor unreliability which are fully integrated into the POPP model.

This chapter starts with a detailed explanation of the POPP extension to model the degree of confidence of sensor reliability and incorporate prior knowledge over the sensor's confidence. The performance of this extension is tested on simulated data, and the result is explained in Section 7.2. Section 7.3 extends the experiment on real world data from previous

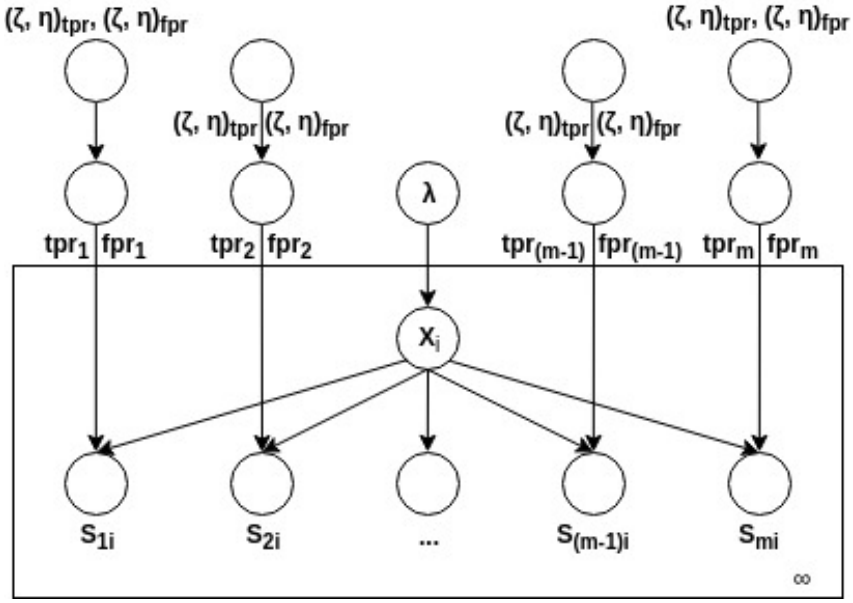

Figure 7.1: Graphical representation of the POPP-Beta.

chapter, and shows the results of the POPP extension on the real world data. Finally, Section 7.4 concludes the contribution, and discusses the limitations of this chapter.

7.1 POPP with Unreliable Sensor Model (POPP-Beta)

The previous chapter has shown how the unreliability of a sensor is incorporated in correctly calculating the posterior over Poisson parameter λ. The sensor model is represented as a confusion matrix which specifies the true positive rate (sensitivity) $\text{tpr}_j = P_j(d = 1 \mid e = 1)$– along with its false negative rate–and the true negative rate (specificity) $\text{fpr}_j = P_j(d = 1 \mid e = 0)$– along with its false positive rate of sensor j . To construct a confusion matrix, one needs to have both detections/predictions and their corresponding actual values. Typically, the collected detections and their corresponding actual counts need expert labelling to pre-process them before they can be used. As one should expect, the POPP model requires the sensor model to be accurate. Otherwise, the posterior over λ will draw incorrect inferences. However, attaining an accurate sensor model requires a lot of training data and this puts a lot of burden on the experts who need to label the data.

Here, the sensor model is transformed into a Bayesian estimation problem where each

element in the confusion matrix, (true positive rate (tpr) and false positive rate (fpr)), follows a beta distribution. Tpr is chosen to represent sensitivity, and fpr is chosen to represent specificity. Beta distributions are chosen for tpr and fpr because Beta distributions act as a conjugate to Binomial distributions and provide a family of prior probability distributions for the parameter of a binomial distribution. The beta-binomial conjugacy leads to an analytically tractable compound distribution called the beta-binomial distribution, where the p parameter in the binomial distribution $B(d \mid x, p)$ is randomly drawn from a beta distribution $Be(p \mid \zeta, \eta)$.

$$\begin{aligned} P(d \mid x,\zeta,\eta) &= \int_{p=0}^{1} P(d \mid x,p)\ P(p \mid \zeta,\eta)\ dp \\ &= \int_{p=0}^{1} B(d \mid x,p)\ Be(p \mid \zeta,\eta)\ dp \\ &= \int_{p=0}^{1} \binom{x}{d} p^{d}(1-p)^{(x-d)}\ \frac{p^{(\zeta-1)}(1-p)^{(\eta-1)}}{\pi(\zeta,\eta)} \\ &= \binom{x}{d} \frac{1}{\pi(\zeta,\eta)} \int_{p=0}^{1} p^{(d+\zeta-1)}(1-p)^{(x-d+\eta-1)}\ dp \\ &= \binom{x}{d} \frac{\pi(d+\zeta, x-d+\eta)}{\pi(\zeta,\eta)} \\ &= BB(d \mid x,\zeta,\eta) \end{aligned} \tag{7.1}$$

As the confusion matrix is now in the form of two beta distributions $Be(\text{tpr} \mid \zeta_{\text{tpr}}, \eta_{\text{tpr}})$ and $Be(\text{fpr} \mid \zeta_{\text{fpr}}, \eta_{\text{fpr}})$, ζ_{tpr} can be thought as the number of true positive detections $\#(d = 1, e = 1)$, η_{tpr} as the number of false negative detections $\#(d = 0, e = 1)$, ζ_{fpr} as the number of false positive detection $\#(d = 1, e = 0)$, and η_{fpr} as the number of true negative detections $\#(d = 0, e = 0)$ that the sensor has made. Given a confusion matrix where the elements of it follow beta density, and beta-binomial distributions which provide an unconditional distribution of d, Equation 6.5 is replaced with:

$$P(s_{ji} \mid x_i) = \sum_{\text{tp}_{ji}=0}^{x_i} BB\Big(\text{tp}_{ji} \mid x_i, \text{tpr}_j, \text{fnr}_j\Big) BB\Big(\text{fp}_{ji} \mid \Delta x_i, \text{fpr}_j, \text{tnr}_j\Big) \tag{7.2}$$

where $s_{ji} = tp_{ji} + fp_{ji}$, $tpr_j = \#_j(d = 1, e = 1)$, $fnr_j = \#_j(d = 0, e = 1)$, $fpr_j = \#_j(d = 1, e = 0)$, $tnr_j = \#_j(d = 0, e = 0)$, and $\Delta x_i = (l - x_i)$.

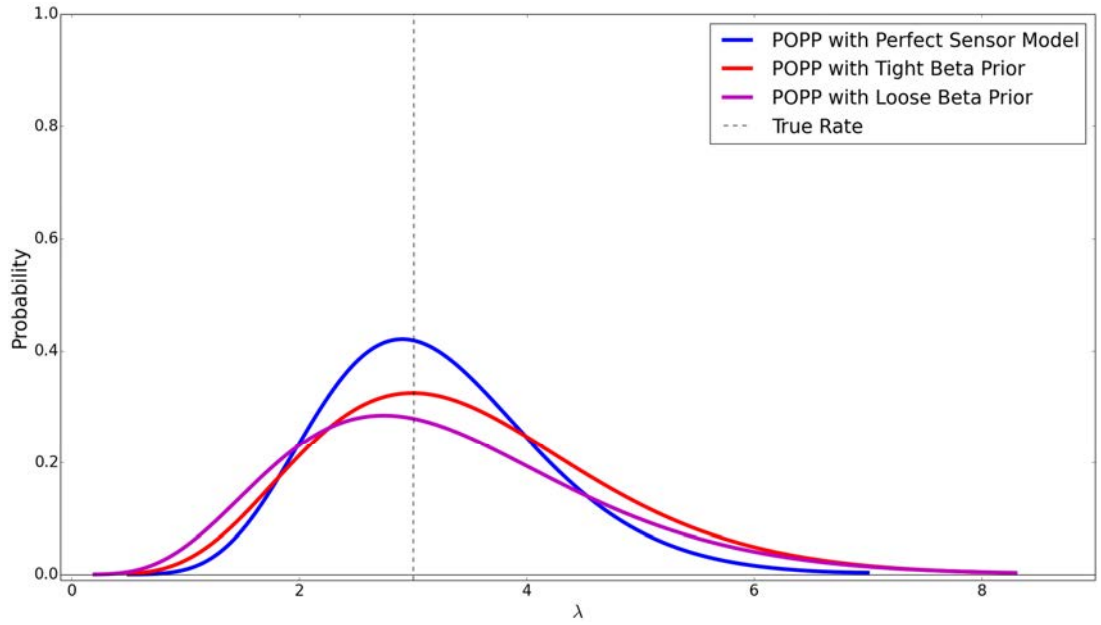

Figure 7.2: The posterior over λ for each filter after the x_{144} sample update with true tpr = 1.0, and true tnr = 0.1. Two shapes of beta prior resulted from the training are presented: loose beta prior (low number of training samples) and tight beta prior (high number of training samples).

With a sensor model which follows beta densities and is fully integrated, as a distribution, in the sensed count likelihood $P(s_{ji} \mid x_i)$ as shown in Equation 7.2, we obtain a graphical model with the structure shown in Figure 7.1. One should note that the difference between the POPP model and this extension, which we call the POPP-Beta model, lies only on altering Equation 6.5 explained in the previous chapter to produce Equation 7.2. However, given little training data for the sensor model, the POPP-Beta model is expected to be more conservative in estimating the posterior $P_G(\lambda \mid \overrightarrow{s})$ over λ than the POPP model.

7.2 Evaluation on Synthetic Data

In the previous chapter, using simulated data, the POPP model establishes that with its associated filters it can outperform the FOPP filter in estimating λ. Here, the same simulated dataset used to evaluate the POPP model in the previous chapter is borrowed. The POPP-Beta model is then examined in estimating λ, and compared its performance against the established POPP model. For this evaluation, the switching filter is chosen as a filter for both

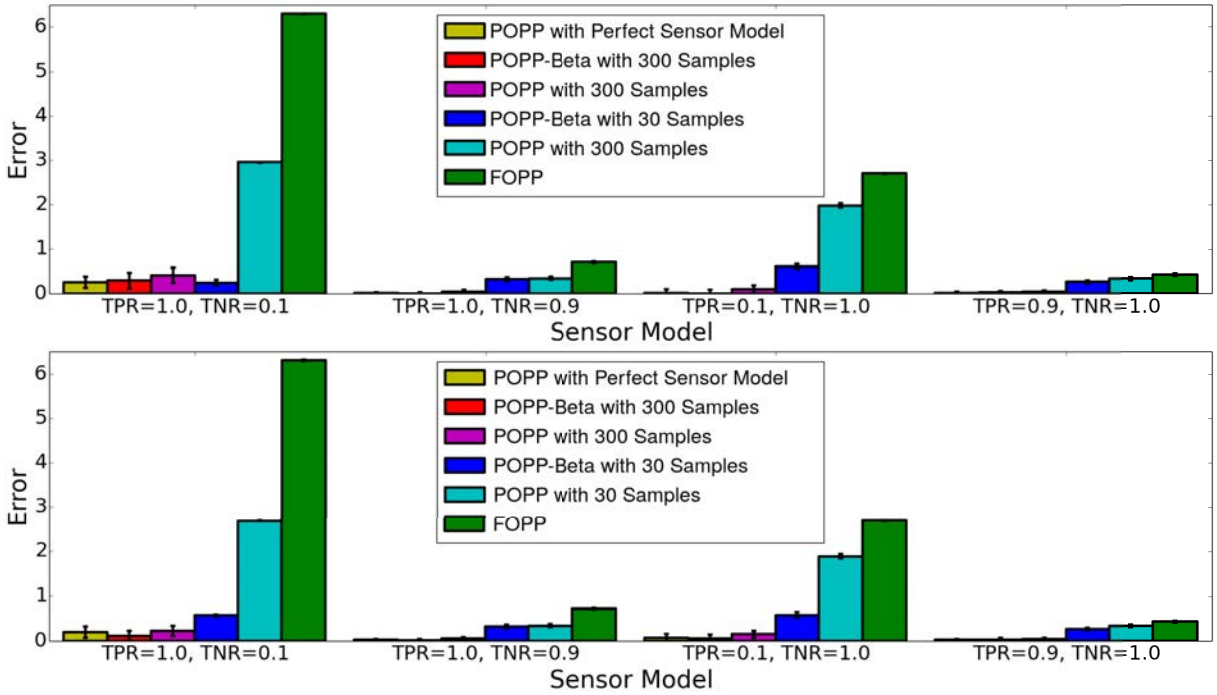

Figure 7.3: The RMSE of posterior estimates of λ for the POPP and the POPP-Beta model with variations on the training data used to build the sensor model. Each trial consisted of a stream of $\overrightarrow{s_1} \ldots \overrightarrow{s_{144}}$ samples. Accuracies of MAP estimates in the top panel, accuracies of expectation of the posterior in the bottom panel. Each data point is an average of 30 trials. Standard errors are shown.

the POPP and the POPP-Beta model. The uncorrected estimate according to the FOPP model was also included in this evaluation.

In each experiment, first, the sensor model was built by sampling simulated counts $x_1, \ldots, x_n$ from a Poisson process $P(x \mid \lambda' = 3)$ together with their corresponding sensor readings. Another set of counts $x_1, \ldots, x_n$ were sampled from the same process to update $P_G(\lambda \mid \overrightarrow{s_i})$ using the switching filter. Figure 7.2 shows an example of the posterior of λ for each filter after the x_{144} sample update.

Two different sample sizes used to build the sensor model are evaluated. A small number of samples builds an erroneous sensor model (for the POPP model) or a loose beta density (for the POPP-Beta model). In contrast, a large number of samples builds a solid sensor model (for the POPP model) or a tight beta density (for the POPP-Beta model). 30 samples were set for the small sample size and 300 samples were set for the large sample size. For a comparison, a POPP model with a perfect sensor model is also included; the correct sensor model is given

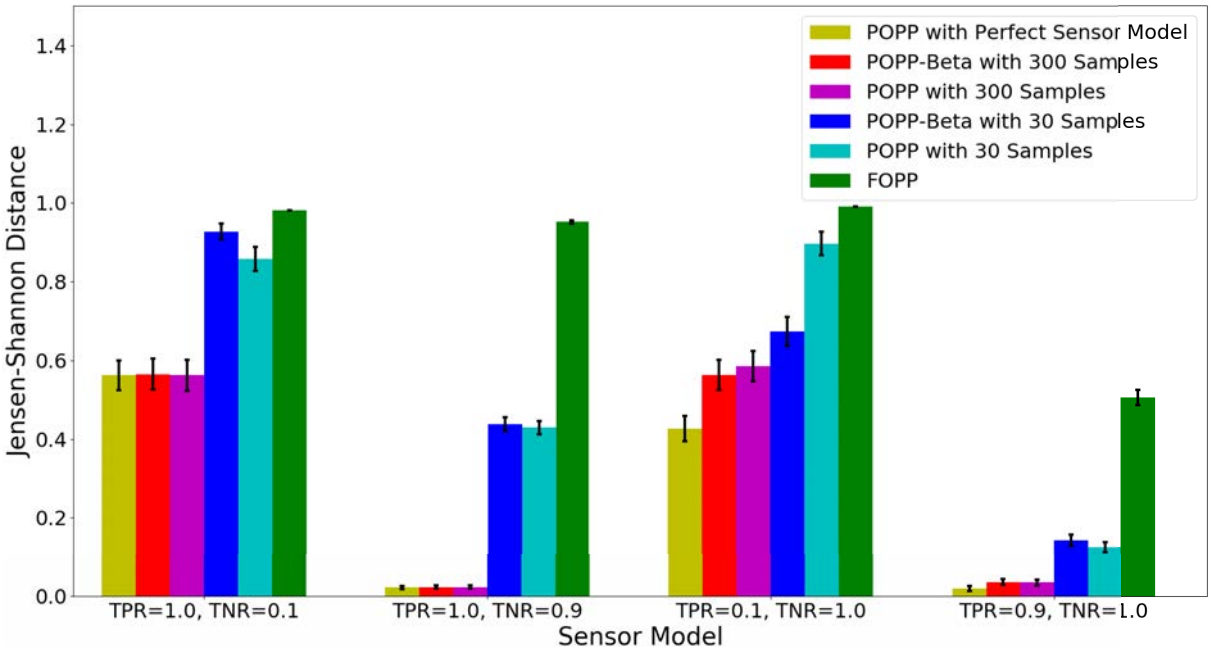

Figure 7.4: The Jensen-Shannon distance of posterior estimates of λ for the POPP and the POPP-Beta model with variations on the training data used to build the sensor model. Each trial consisted of a stream of $\overrightarrow{s_1} \ldots \overrightarrow{s_{144}}$ samples. Each data point is an average of 30 trials. Standard errors are shown.

without prior training.

The POPP model and the POPP-Beta model are evaluated on four different levels of sensor unreliability:

- True positive rate: 0.1, true negative rate: 1.0;
- True positive rate: 0.9, true negative rate: 1.0;
- True positive rate: 1.0, true negative rate: 0.1;
- True positive rate: 1.0, true negative rate: 0.9

with the reasoning that any other level of sensor unreliability lies between these four.

The performance of the POPP-Beta model was assessed similarly as in the previous chapter by comparing the RMSE (the MAP hypothesis (mode) and the expectation (mean)) and the Jensen-Shannon distance of the two estimators over λ to the true λ'. The result of the assessment is shown in Figures 7.3 and 7.4. Figure 7.3 shows that the POPP-Beta model performed better than the standard POPP model for both a low number of training samples

and a high number of training samples for building the sensor model. Figure 7.4 displays a different perspective. It shows that the performance of the POPP-Beta filter was comparable to that of the POPP filter. However, the special case of when the true positive rate is low, the POPP-Beta filter outperformed the POPP filter.

7.3 Evaluation on A Real World Dataset

Having shown the performance of POPP-Beta against the POPP model on simulated data, the evaluation of the performance of POPP-Beta model moves on a real world data set. Similar to the evaluation on simulated data, the performance of the POPP-Beta model was compared againsts the standard POPP model and the FOPP model.

The dataset, which has been explained in the previous chapter, is a collection of counts over time from leg detector (LD), upper body detector (UBD), and change detector (CD) gathered by a mobile robot which observes and counts the number of people passing by from 10am-8pm on each day. For this evaluation, the 69-day deployment dataset was used. One should remember that the detections are still temporally and spatially scattered and they are not uniformly distributed across space. The detections are organised according to time/date and the spatial region where each detection was made.

Both the POPP-Beta model and the standard POPP model rely on sensor models which must be calculated from sensor counts and true counts. Four fold cross-validation to separate the training and testing data were performed, i.e., 12-days of data were used as a training set for a sensor model, and the remaining 36-days of data were used as a test set. Both the POPP-Beta and the standard POPP model used the switching filter as its filter to do the inferences from the sensor counts.

Similar to the previous chapter, periodic Poisson processes are included as a showcase comparison whether the non-homogeneous Poisson processes, in which the periodic Poisson processes are categorised into, are benefited by the POPP filters.

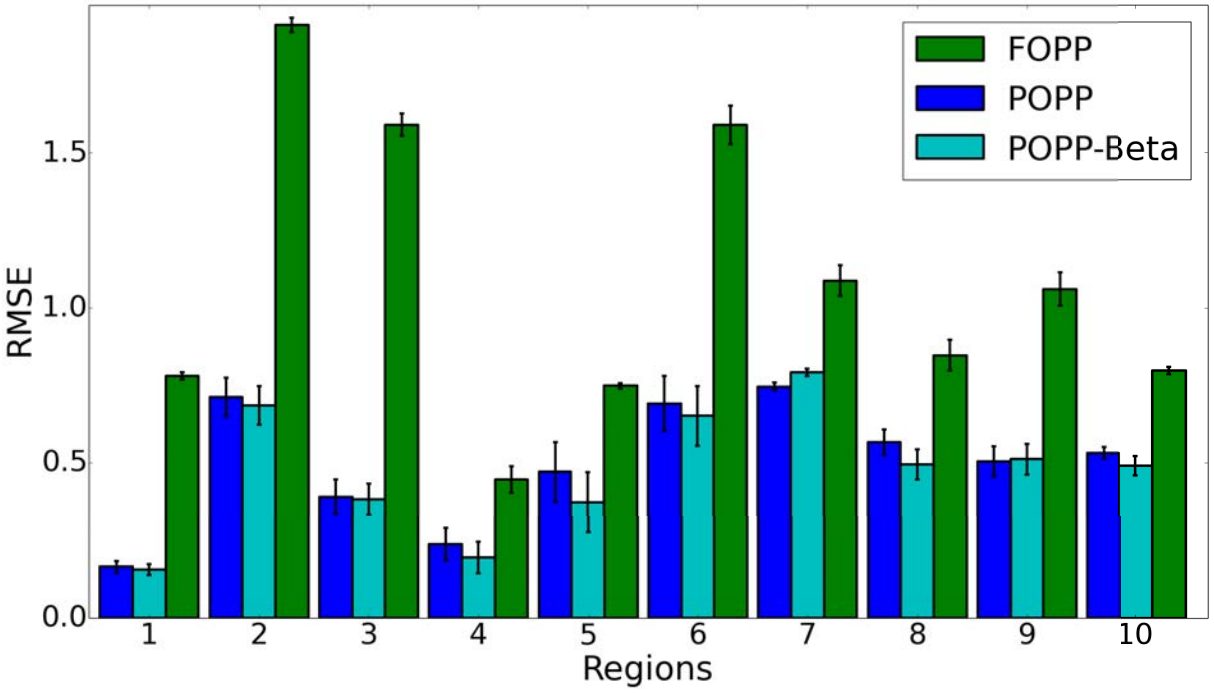

Figure 7.5: The RMSE of the FOPP, POPP, and POPP-Beta estimators of λ as it varies across areas (regions) of the environment. Standard error is shown.

7.3.1 The POPP-Beta models on Homogeneous Poisson Processes

This section starts with a case on homogeneous Poisson processes. The interval time for the Poisson is set to 10 minutes, and the true counts and the detections made by each detector were recorded. The true λ of the Poisson distribution for each region is estimated by running a FOPP filter on the true counts. The uncorrected estimate λ according to the FOPP model was estimated only from the change detector count data.

Both point estimate predictions based on the MAP estimate and their approximation distributions of each model were then compared to the true λ and its distribution using RMSE and Jensen-Shannon distance on region by region basis.

The results are shown in Figure 7.5 and 7.6. In terms of the metric distance, the POPP-Beta filter, in general, produced better estimates than the POPP filter. It is, by extension, more accurate than the FOPP filter. Recall that the distribution of detections is not uniform across areas, some areas (like area 4, 5, 7, and 9) have fewer data compared to other areas such that the underlying sensor models for both the POPP-Beta and the POPP are relatively unreliable.

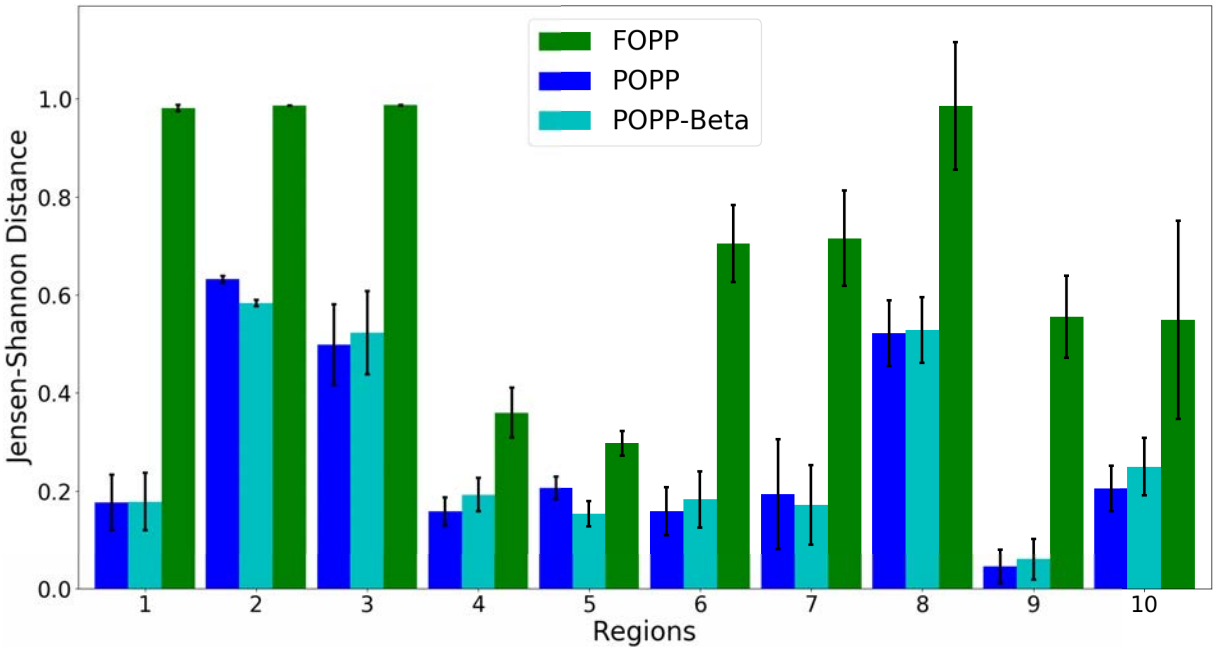

Figure 7.6: The Jensen-Shannon distance of the FOPP, the POPP, and the POPP-Beta model distributions of λ as it varies across areas (regions) of the environment. Standard error is shown.

In area 4 and 5, the point estimates of the POPP-Beta filter is a little bit more accurate than the ones of the POPP filter. In areas 7 and 9, however, the standard POPP filter managed to slightly outperform the POPP-Beta filter. This is attributed to the sensor model which follows beta densities has a wide distribution (loose beta prior) due to limited learning data. The opposite was true for some areas such as area 1, 2, and 3, where there are plenty of data to construct the sensor models resulting a constant better gain in RMSE over the POPP estimator.

On the other hand, the Jensen-Shannon distance shows that the POPP-Beta seemed to produce slightly different distributions than the POPP, i.e. having higher Jensen-Shannon distance than the POPP model. This can be explained by Figure 7.2. The POPP-Beta drives the estimate λ in a more conservative way resulting an estimate distribution that is typically wider than the one produced by the POPP.

As a side evaluation, how the POPP-Beta evolved with time was also recorded, again in terms of both RMSE and Jensen-Shannon distance, since it is expected to gradually get closer to the true λ. Figures 7.7 and 7.8 show that as time passes the performance of the POPP-Beta estimator steadily becomes better and outperforms the standard POPP model.

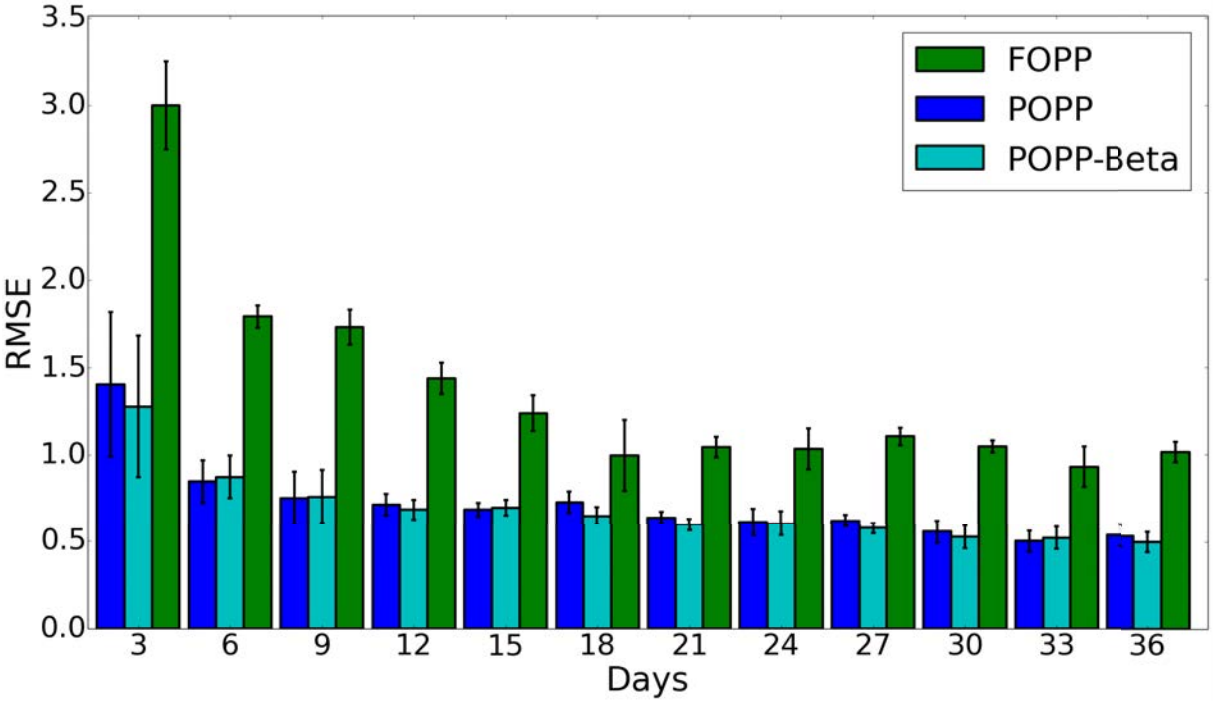

Figure 7.7: The RMSE evolution from day 3 to day 36 with 3 day interval, averaged across all regions. Standard error is shown.

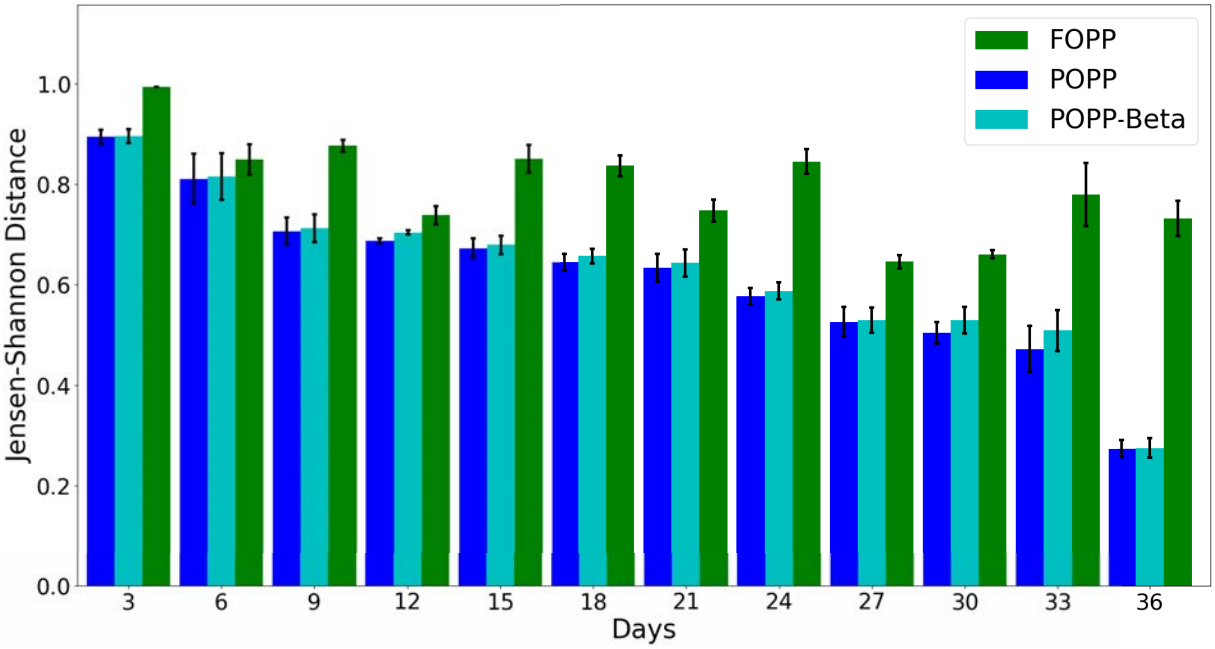

Figure 7.8: The Jensen-Shannon distance evolution of the FOPP, the POPP, and the POPP-Beta model distributions of λ from day 3 to day 36 with 3 day interval, averaged across all regions. Standard error is shown.

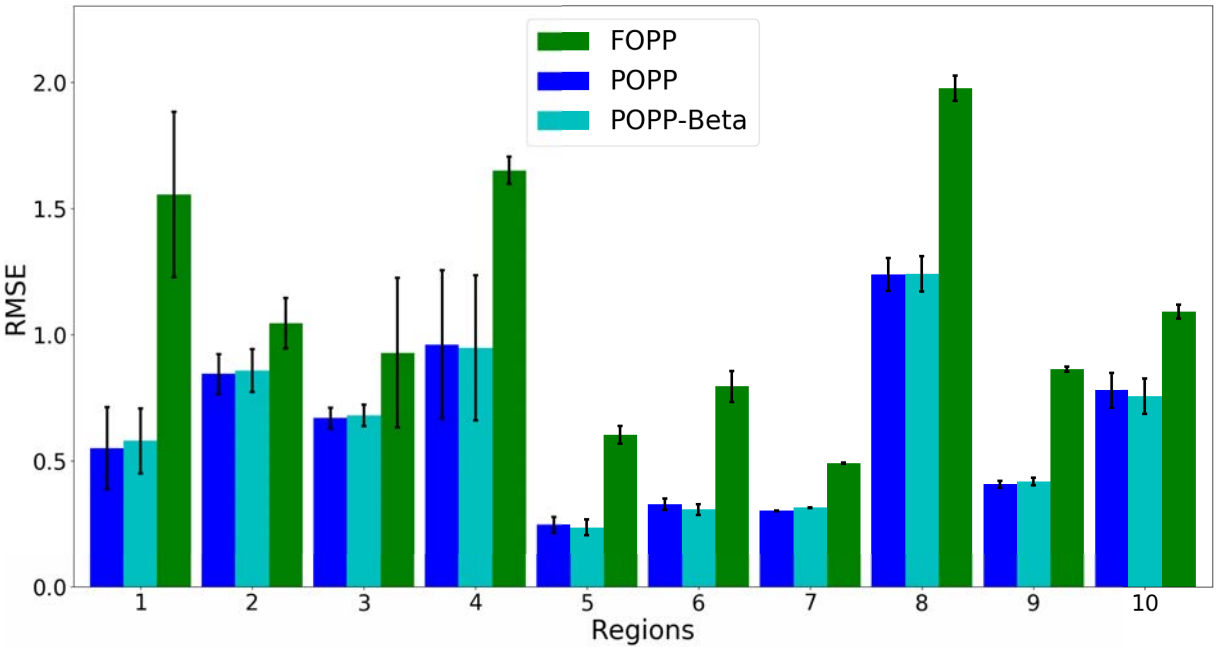

Figure 7.9: The RMSE of the FOPP, POPP, and POPP-Beta estimators of $\lambda(t_i, t_j)$ as it varies across areas (regions) of the environment. Standard error is shown.

7.3.2 The POPP-Beta models on Periodic Poisson Processes

For this evaluation, the underlying process on each region was then assumed to be a periodic Poisson process in which the imposed single periodicity is a one-day cycle, i.e. $\lambda(t_i, t_j) = \lambda(t_{i+\Delta}, t_{j+\Delta})$ with $\Delta = 24 * 60$(minutes). The true $\lambda(t_i, t_j)$ of the periodic Poisson processes for each region is estimated by running a FOPP filter on the true counts. The uncorrected estimate $\lambda(t_i, t_j)$ according to the FOPP model was estimated only from the change detector count data.

The MAP estimate of each model was then compared to the true $\lambda(t_i, t_j)$ using RMSE, whereas the approximated distributions of each model was compared to the true continuous distribution of $\lambda(t_i, t_j)$ using Jensen-Shannon distance. The comparison was done for all areas within the patrol space.

Figures 7.9 and 7.10 show that the POPP-Beta filter gave comparable performance to the standard POPP filter which came better compared to the FOPP filter. Same arguments concluding the performance of the POPP filter can be applied to the POPP-Beta model since the POPP-Beta model drives the $\lambda'(t_i, t_j)$ to the true $\lambda(t_i, t_j)$ in a more conservative way than

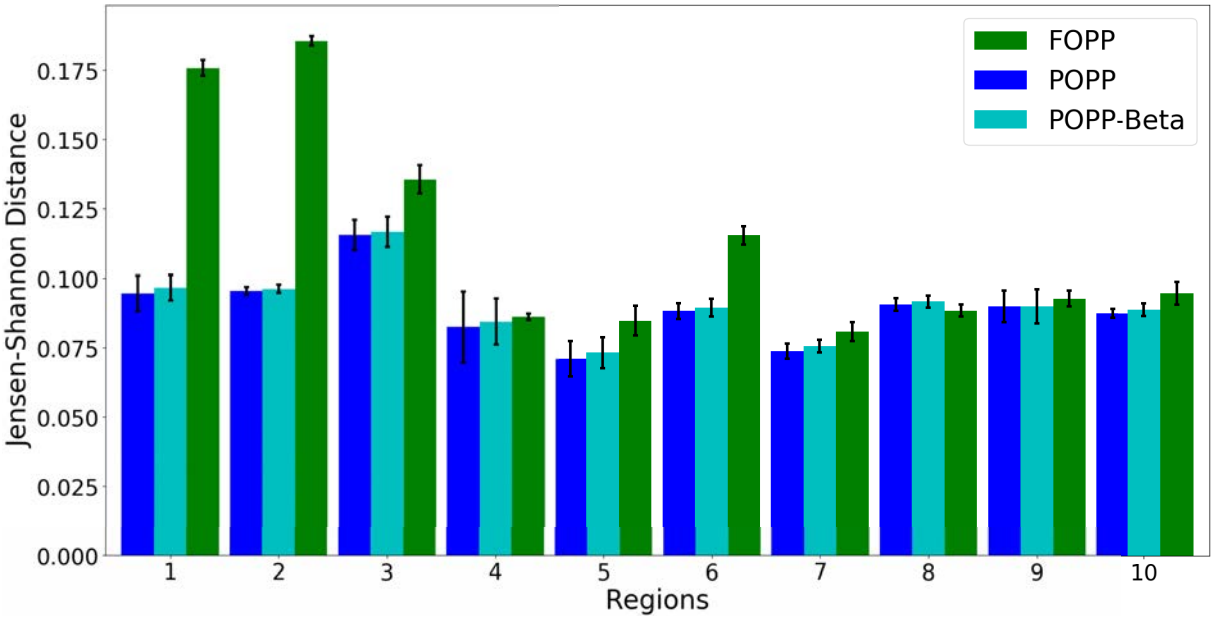

Figure 7.10: The Jensen-Shannon distance of the FOPP, POPP, and POPP-Beta model distributions of $\lambda(t_i, t_j)$ as it varies across areas (regions) of the environment. Standard error is shown.

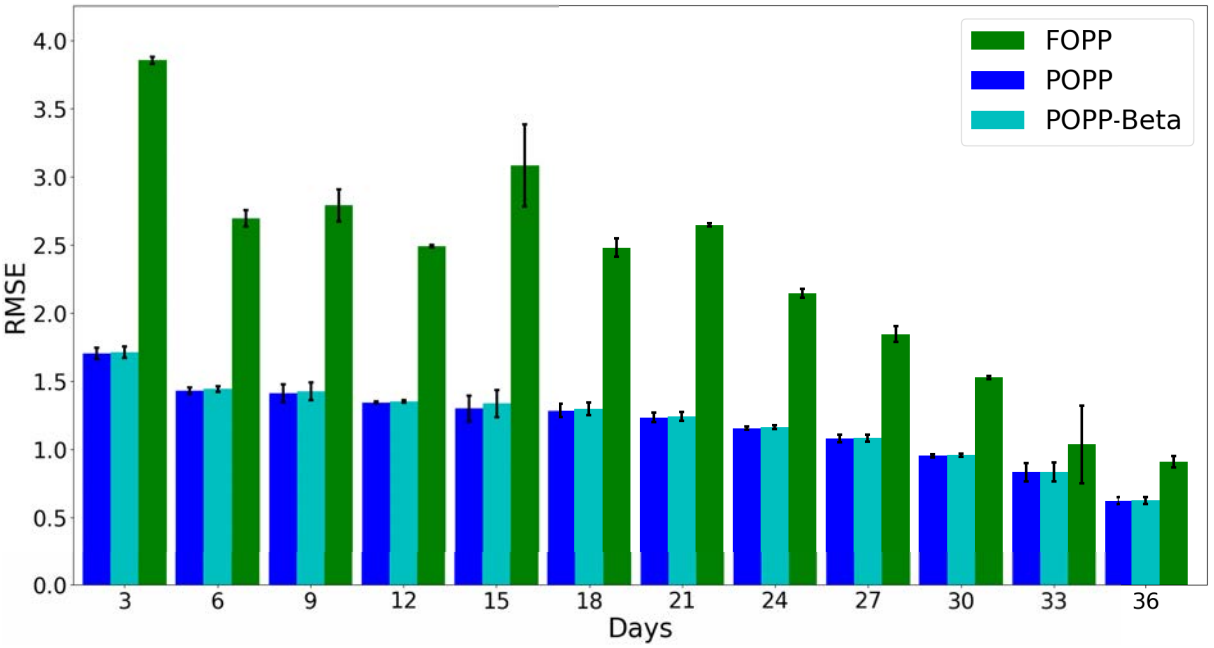

Figure 7.11: The RMSE evolution of periodic Poisson processes with POPP, POPP-Beta, and FOPP filters from day 3 to day 36, averaged across all regions. Standard error is shown.

the standard POPP filter.

How the POPP-Beta evolved with time was also recorded as a side evaluation. Figures 7.11 and 7.12 show the evolution of each filter overtime in terms of RMSE and Jensen-

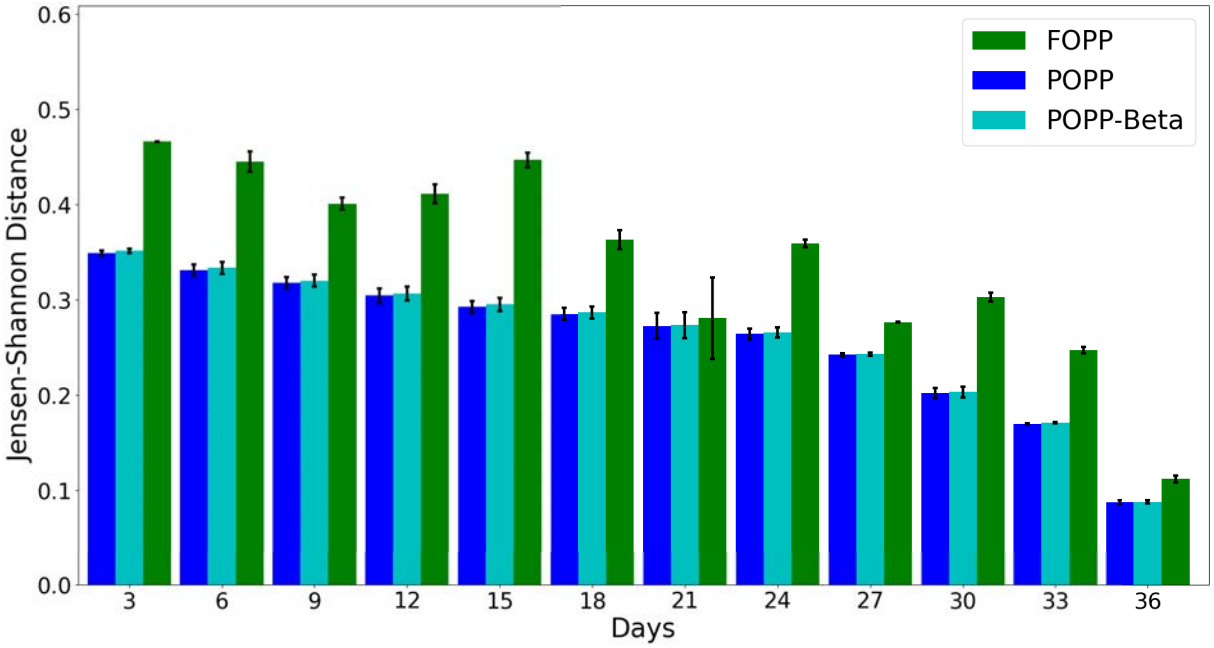

Figure 7.12: The Jensen-Shannon distance evolution of the FOPP, the POPP, and the POPP-Beta filters in periodic Poisson processes from day 3 to day 36 in a 3-day interval, averaged across all regions. Standard error is shown.

Shannon distance across regions. As time passed by, the POPP-Beta filter increased its accuracy in estimating $\lambda(t_i, t_j)$. Similar to what had shown in 7.7 and 7.8, the POPP-Beta filter has similar accuracy to the one shown by the POPP filter. This translates to the POPP-Beta filter together with the POPP filter being more accurate than the FOPP filter.

7.4 Discussion

This chapter focused on the extension of the partially observable Poisson process (POPP) to remove the POPP's requirement to have a precise sensor model to be known in advanced. It is motivated by the fact that attaining a precise sensor model requires a lot of training data and expert interventions.

To achieve this, a fixed point sensor model is transformed into a Bayesian estimation where the TPR and TNR of the sensor model follow beta distributions. Due to the nature of conjugacy between beta distribution and the binomial distribution in Equation 6.5, an analytical closed form beta-binomial distribution was proposed to replace the binomial

distribution in Equation 6.5.

The experiments on simulated data showed that the extension–the POPP-Beta–outperformed a standard POPP. It did, by extension, outperform a standard Bayes' filter for the FOPP model in estimating the parameter λ of a Poisson process. On real world data, similar results are achieved in the case of homogeneous Poisson processes that show how the POPP-Beta model outperformed the standard POPP model. Similar to the POPP model, the posterior of the POPP-Beta model is able to envelop the true λ. However, the POPP-Beta model was not able to tackle the problem that the POPP model faced when it comes to the case of having limited sample data both to build the sensor model and to estimate the true λ. In the periodic Poisson process cases where sample data were limited, the POPP-Beta model gave a similar performance to the POPP model which, in some cases, is outperformed by the FOPP model.

The POPP, which provides a way to deal with counts coming from unreliable counting devices, is extended with the POPP-Beta to deal with unreliable sensor models. This nicely tackles one of the two assumptions that the POPP has. This leaves us with another assumption which is the assumption that counter failures are conditionally independent from one to another given the true count.

CHAPTER 8

PARTIALLY OBSERVABLE POISSON PROCESS: NON-INDEPENDENT SENSOR FAILURES

As detailed in Chapter 6, the POPP model relies on two different assumptions to be able to correctly estimate counts made by multiple unreliable counting devices. The previous chapter studied a way to remove one of the POPP assumptions, which is the assumption of knowing precisely the unreliability of a sensor. In other words, the sensor model must be known in advanced. There is one assumption unaddressed of which is:

- POPP assumes an independent sensor assumption where a failure in correctly detecting an event of a sensor is uncorrelated to one another.

This chapter focuses on addressing this remaining assumption. A joint sensor model is introduced and, with this notion, Equation 6.3 is updated. The concept is fully integrated into the original POPP model.

First, an extension to the POPP model to model the joint sensor is introduced. The performance of this extension is tested on simulated data, and the result is explained in Section 8.2. Section 8.3 extends the experiment on real world data from the previous chapter and shows the results of the POPP extension on this. Finally, Section 8.4 concludes the chapter, and discusses the remaining limitations.

8.1 Non-Independent Partially Observable Poisson Process (N-POPP)

Recall that Equation 6.3 is defined under the assumption that each sensor's count is conditionally independent from all the others given the true count. This assumption ignores the correlations between sensors. Consequently, $P(\overrightarrow{s_i} \mid x_i)$ is defined as a simple product across the sensors.

$$P(\overrightarrow{s_i} \mid x_i) = \prod_{j=1}^{m} P(s_{ji} \mid x_i) \tag{8.1}$$

with $1 \leq i \leq n$, $\overrightarrow{s_i} = (s_{1i}, \ldots, s_{mi})$ and $P(s_{ji} \mid x_i)$ defined in Equation 6.5.

By removing the assumption, the model gives the possibility that there are correlations among sensors. A modification of the model is necessary to accommodate this relaxation because:

1. Equation 8.1 holds only for the independent sensor assumption, and

2. Equation 6.5 becomes inappropriate because it only involves an independent sensor model, represented by true positive detection (tpr_j) $P_j(d_k = 1 \mid e_k = 1)$ and the false positive detection (fpr_j) $P_j(d_k = 1 \mid e_k = 0)$, for each individual sensor j.

Similar to the derivation of Equation 6.5, an arbitrarily close approximation to the probability $P(\overrightarrow{s_i} \mid x_i)$ is defined by assuming there exists a small enough finite subinterval of length δ for which the probability of more than one event occurring is less than some small value ϵ. Once again, the interval (t_1, t_2), where the true count x_i occurred, is split into l smaller

subintervals $I_1, \ldots, I_l$ of equal size with the condition that $l > \lambda$. Consequently, for the event e_k, the whole interval $(t_1, t_2) = I_1, \ldots, I_l$ follows the Bernoulli distribution, where the k^{th} trial corresponds to whether an event e_k happens with probability λ / l at the subinterval I_k. For the detections, however, the whole interval $(t_1, t_2) = I_1, \ldots, I_l$ does not follow the Bernoulli distribution anymore, but rather follows the categorical distribution, where the k^{th} trial corresponds to whether a particular combination of binary detections $d_{1k}, \ldots, d_{mk}$ happens in subinterval I_k. Hence, instead of having an independent sensor model for each sensor j:

$$tpr_j = P_j(d_k = 1 \mid e_k = 1), fpr_j = P_j(d_k = 1 \mid e_k = 0)$$

with $i \leq j \leq m$, a joint sensor model is needed:

$$P_{jnt}(d_{1k}, \ldots, d_{mk} \mid e_k) \tag{8.2}$$

with d_{jk} being a detection by sensor j at subinterval I_k and $d_{jk}, e_k \in [0, 1]$. The variation of $P_{jnt}(d_{1k}, \ldots, d_{mk} \mid e_k)$ grows by a factor of 2 with the number of sensors involved. Regardless the number of sensors, $P_{jnt}(d_{1k}, \ldots, d_{mk} \mid e_k)$ can be split into two:

1. $P_{jnt}(d_{1k}, \ldots, d_{mk} \mid e_k = 1)$ for the probability of a combination of binary detections $d_{1k}, \ldots, d_{mk}$ given that the event e_k happened at subinterval k. Let $\mathscr{E}^+$ be a set of probabilities $P_{jnt}(d_{1k}, \ldots, d_{mk} \mid e_k = 1)$, then $\mathscr{E}^+$ includes all possible combinations of binary detections $d_{1k}, \ldots, d_{mk}$ given $e_k = 1$. Let $E_0^+ = P_{jnt}(d_{1k} = 0, \ldots, d_{mk} = 0 \mid e_k = 1), \ldots, E_{(m^2)-1}^+ = P_{jnt}(d_{1k} = 1, \ldots, d_{mk} = 1 \mid e_k = 1)$, then $\mathscr{E}^+ = \left\{E_0^+, \ldots, E_{(m^2)-1}^+\right\}$ is called the *true joint positive rate*.

2. $P_{jnt}(d_{1k}, \ldots, d_{mk} \mid e_k = 0)$ for the probability of a combination of binary detections $d_{1k}, \ldots, d_{mk}$ given that the event e_k did not happened at subinterval k. Let $\mathscr{E}^-$ be a set of probabilities $P_{jnt}(d_{1k}, \ldots, d_{mk} \mid e_k = 0)$, then $\mathscr{E}^-$ includes all possible combinations of binary detections $d_{1k}, \ldots, d_{mk}$ given $e_k = 0$. Let $E_0^- = P_{jnt}(d_{1k} = 0, \ldots, d_{mk} = 0 \mid e_k =$

$1), \ldots, E^{-}_{(m^2)-1} = P_{jnt}(d_{1k} = 1, \ldots, d_{mk} = 1 \mid e_k = 1)$, then $\mathscr{E}^{-} = \{E^{-}_0, \ldots, E^{-}_{(m^2)-1}\}$ is called the *true joint negative rate*.

Recall that the set of detections $\overrightarrow{s_i}$ was defined as:

$$\overrightarrow{s_i} = (s_{1i}, \ldots, s_{mi})$$

with $s_{ji} \in \mathbb{N}$. s_{ji} is the sensed count of sensor j from the i-th observation. Since the joint sensor model is defined under a combination of binary detections of sensors, each s_{ji} is split into l subintervals such that in each sub interval I_k there is only one detection d_{jk}. If the binary detections from all sensors at subinterval I_k are grouped together, then for the interval i, $\overrightarrow{s_i}$ can be seen as a collection of l groups of binary detections

$$\overrightarrow{s_i} = ((d_{11}, \ldots, d_{m1}), \ldots, (d_{1l}, \ldots, d_{ml})) \tag{8.3}$$

with d_{jk} being a detection by sensor j at subinterval I_k and $d_{jk} \in [0, 1]$.

$\overrightarrow{s_i}$ may contain groups with similar binary detections $d_{jk}, 1 \leq j \leq m, 1 \leq k \leq l$. Let $D_0 = (d_{1k}, \ldots, d_{mk}) = (0, \ldots, 0)$ (all $d_{jk} = 0$), $D_1 = (d_{1k}, \ldots, d_{mk}) = (0, \ldots, 1)$ (all $d_{jk} = 0$ except d_{mk}), …, $D_{(m^2)-1} = (d_{1k}, \ldots, d_{mk}) = (1, \ldots, 1)$. It is straight forward to say that for each $\overrightarrow{s_i}$, each group $(d_{1k}, \ldots, d_{mk}) \in \{D_0, \ldots, D_{(m^2)-1}\}$ for $1 \leq k \leq l$. With this, $\overrightarrow{s_i}$ definition from 8.3 can be redefined as

$$\overrightarrow{s_i} = (D_0 = g_0, \ldots, D_{(m^2)-1} = g_{(m^2)-1}) \tag{8.4}$$

with $g_0, \ldots, g_{(m^2)-1} \in \mathbb{N}$. With the definition in 8.4, $\overrightarrow{s_i}$ groups a detection of each sensor from the same subinterval as a joint detection, represented by $D_q, 0 \leq q \leq (m^2) - 1$, and displays the number of the appearance of the joint detection, represented by $g_q, 0 \leq q \leq (m^2) - 1$, within the interval (t_1, t_2). We further define:

$$\begin{aligned} \| \overrightarrow{s_i} \| &= \| (D_0 = g_0, \ldots, D_{(m^2)-1} = g_{(m^2)-1}) \| \\ &= \sum_{h=0}^{(m^2)-1} g_h \end{aligned} \tag{8.5}$$

with $\overrightarrow{s_i}$ as in definition 8.4, and $g_0, \ldots, g_{(m^2)-1} \in \mathbb{N}$. One should see clearly that $\| \overrightarrow{s_i} \| = l$ given that the interval is split into l subintervals.

The probability of $\overrightarrow{s_i}$ given x_i events happened is now the aggregate of joint detections given the positive event $P_{jnt}(d_{1k}, \ldots, d_{mk} \mid e_k = 1)$ in x_i sub-intervals, and joint detections given the negative event $P_{jnt}(d_{1k}, \ldots, d_{mk} \mid e_k = 0)$ in $l - x_i$ sub-intervals. Given the probability of $P_{jnt}(d_{1k}, \ldots, d_{mk} \mid e_k)$ for all possible combinations of d_{jk} and e_k, equation 8.1 is redefined as:

$$P(\overrightarrow{s_i} \mid x_i) = \sum_{\overrightarrow{s^+} \subseteq \overrightarrow{s_i}} Multi(\overrightarrow{s^+} \mid x_i, \mathscr{E}^+)\ Multi(\overrightarrow{s^-} \mid (l - x_i), \mathscr{E}^-) \tag{8.6}$$

with $\| \overrightarrow{s^+} \| = x_i$, $\| \overrightarrow{s_i} \| = \| \overrightarrow{s^+} \| + \| \overrightarrow{s^-} \|$.

One should note that Equation 8.6 mainly aims for estimating the sensed count likelihood on multiple sensors to utilise the correlation among sensors in detections. If there is only one sensor counting events, then the POPP model with the independent sensor assumption (utilising equation 8.1) is more computationally efficient due to Equation 6.5.

8.2 Evaluation on Synthetic Data

Chapter 6 has shown how the POPP model and its associated filters outperform the FOPP filter in accurately estimating λ. Here, an evaluation and a comparison of the N-POPP model to the POPP model are conducted with two imaginary unreliable sensors and simulated datasets. The switching filter is chosen as a filter for both the POPP and the N-POPP model for this evaluation. As the basis, a comparison against the uncorrected estimate according to the FOPP model is also included.

In each experiment, the simulated sensors have a predefined joint sensor model which is used to produce the sensor readings $\overrightarrow{s_1} \ldots \overrightarrow{s_n}$ given the true counts $x_1, \ldots, x_n$ sampled from a Poisson process $P(x \mid \lambda' = 3)$. The estimated joint sensor model is then built based on the sampled counts $x_1, \ldots, x_n$ and the corresponding sensor readings $\overrightarrow{s_1} \ldots \overrightarrow{s_n}$. Then another set of

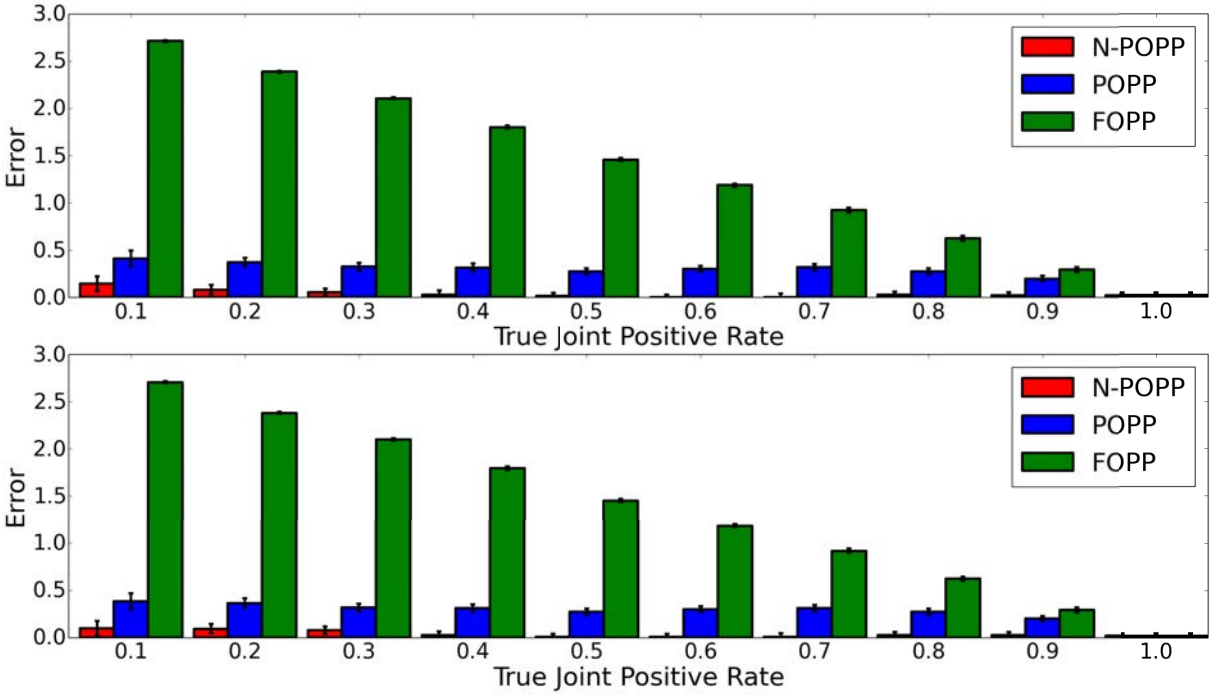

Figure 8.1: The RMSE of posterior estimates of λ for the N-POPP and the POPP model with variation on $\mathscr{E}^+$ with positive correlation between two sensors. Accuracies of MAP estimates in the top panel, accuracies of expectation of the posterior in the bottom panel. Each data point is an average of 30 trials. Standard errors are shown.

counts $x_1, \ldots, x_{144}$ is sampled from the same process. These counts were then fed once again to a simulated sensor producing sensor readings $\overrightarrow{s_1} \ldots \overrightarrow{s_{144}}$. A recursive update, then, takes place on $P_G(\lambda \mid \overrightarrow{s_i})$ using the estimated joint sensor model and the switching filter.

The performance of the N-POPP filter, the standard POPP filter, and the standard FOPP filter were assessed by comparing the RMSE of two estimators: the MAP hypothesis (mode) and the expectation (mean) of each posterior over λ to the true λ'. The referenced joint sensor model $P_{jnt}(d_{1k}, \ldots, d_{mk} \mid e_k)$ is varied and the evaluation of the performance is split into two parts. First evaluation is the evaluation of the performance of the N-POPP filter and its corresponding POPP filter by varying the true joint positive rate $\mathscr{E}^+$ (similar to varying sensor sensitivity in the POPP model) while fixing the true joint negative rate $\mathscr{E}^-$. The variation falls into three categories:

1. The variation on the true joint positive rate $\mathscr{E}^+$ where two sensors are positively correlated in positive detections. This is shown by $P_{jnt}(d_1 = 1, d_2 = 1 \mid e = 1) > 0.0$, and

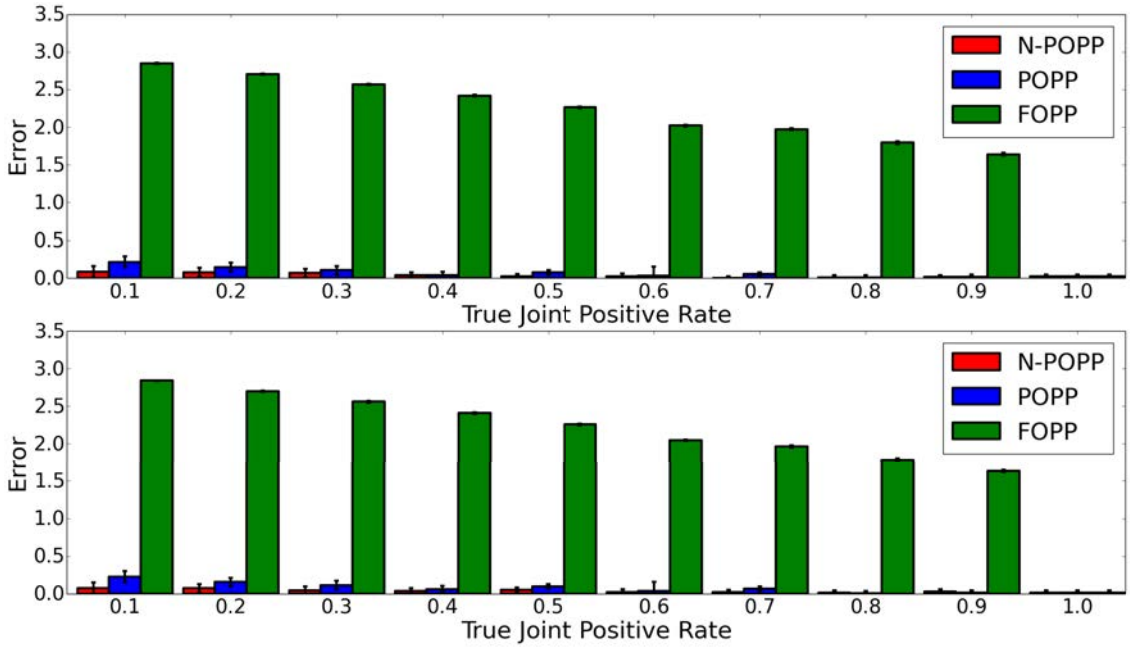

Figure 8.2: The RMSE of posterior estimates of λ for the N-POPP and the POPP model with variation on $\mathscr{E}^+$ with negative correlation between two sensors. Accuracies of MAP estimates in the top panel, accuracies of expectation of the posterior in the bottom panel. Each data point is an average of 30 trials. Standard errors are shown.

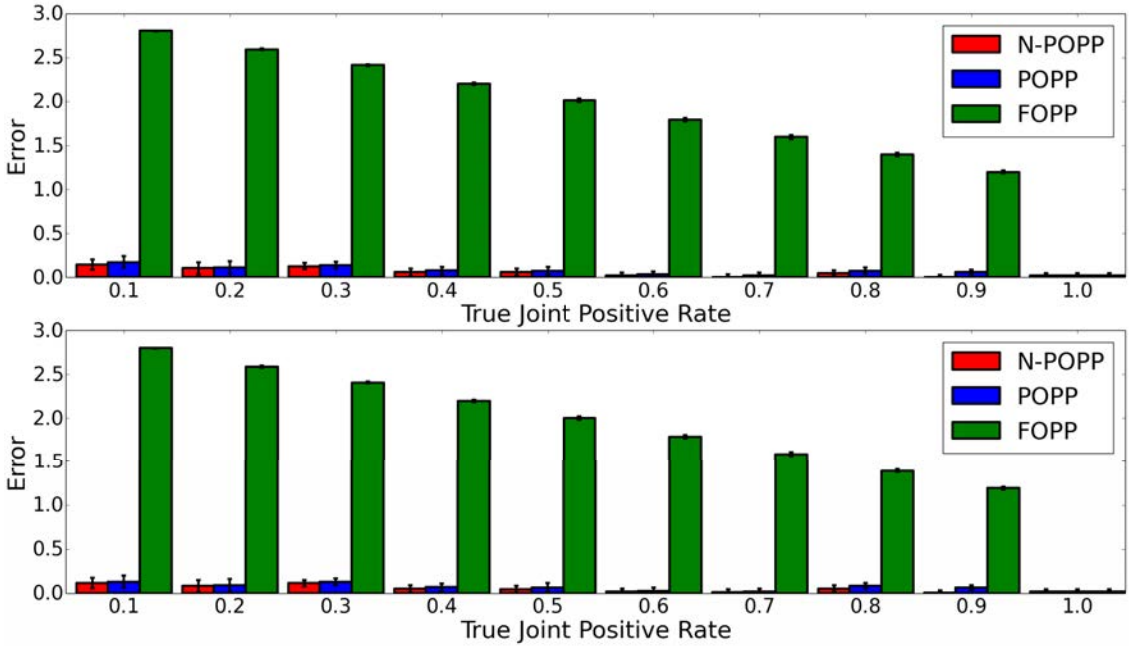

Figure 8.3: The RMSE of posterior estimates of λ for the N-POPP and the POPP model with variation on $\mathscr{E}^+$ with no correlation between two sensors. Accuracies of MAP estimates in the top panel, accuracies of expectation of the posterior in the bottom panel. Each data point is an average of 30 trials. Standard errors are shown.

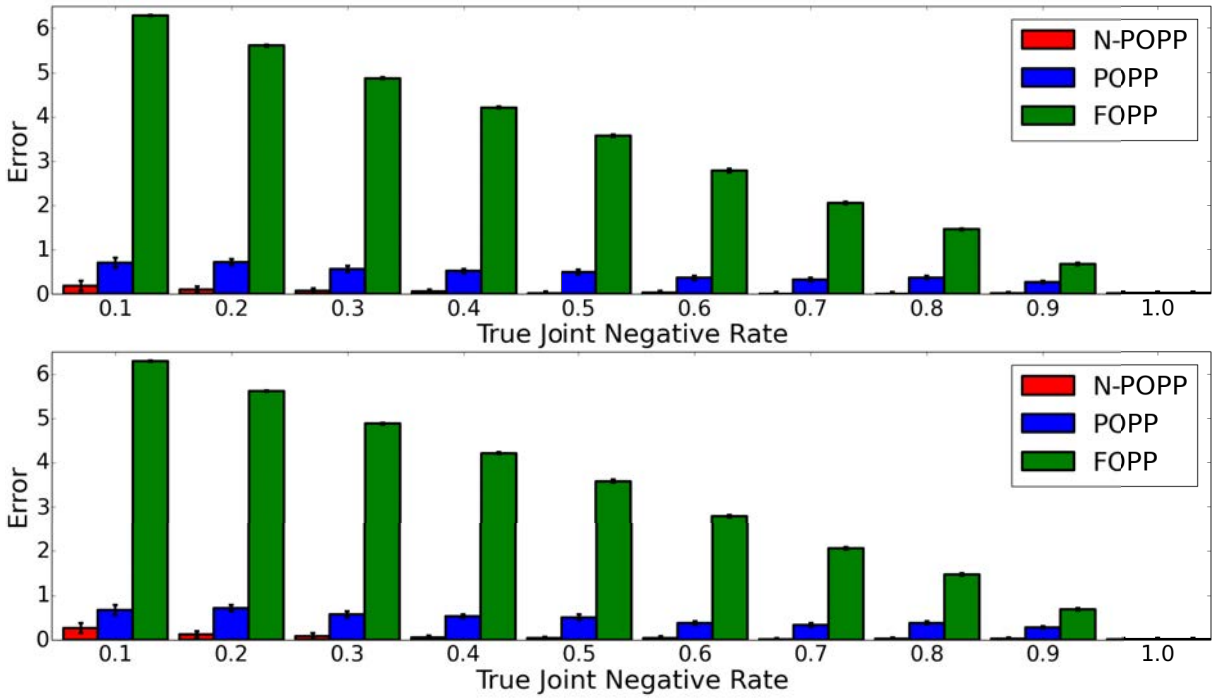

Figure 8.4: The RMSE of posterior estimates of λ for the N-POPP and the POPP model with variation on $\mathscr{E}^-$ with positive correlation between two sensors. Accuracies of MAP estimates in the top panel, accuracies of expectation of the posterior in the bottom panel. Each data point is an average of 30 trials. Standard errors are shown.

$P_{jnt}(d_1 = 0, d_2 = 1 \mid e = 1) = P_{jnt}(d_1 = 1, d_2 = 0 \mid e = 1) = 0.0$. The performance of this variation is shown in Figure 8.1.

2. The variation on the true joint positive rate $\mathscr{E}^+$ where two sensors are negatively correlated in positive detections. This is shown by $P_{jnt}(d_1 = 0, d_2 = 1 \mid e = 1) > 0.0$, $P_{jnt}(d_1 = 1, d_2 = 0 \mid e = 1) > 0.0$, and $P_{jnt}(d_1 = 1, d_2 = 1 \mid e = 1) = 0.0$. The performance of this variation is shown in Figure 8.2.

3. The variation on the true joint positive rate $\mathscr{E}^+$ where two sensors are uncorrelated in positive detections (similar to the POPP model with an independent assumption). This is shown by $P_{jnt}(d_1 = 0, d_2 = 1 \mid e = 1) > 0.0$, $P_{jnt}(d_1 = 1, d_2 = 0 \mid e = 1) > 0.0$, and $P_{jnt}(d_1 = 1, d_2 = 1 \mid e = 1) > 0.0$. The performance of this variation is shown in Figure 8.3.

The second evaluation is the evaluation of the performance of the N-POPP filter and its corresponding POPP filter by varying the true joint negative rate $\mathscr{E}^-$ (similar to varying sensor

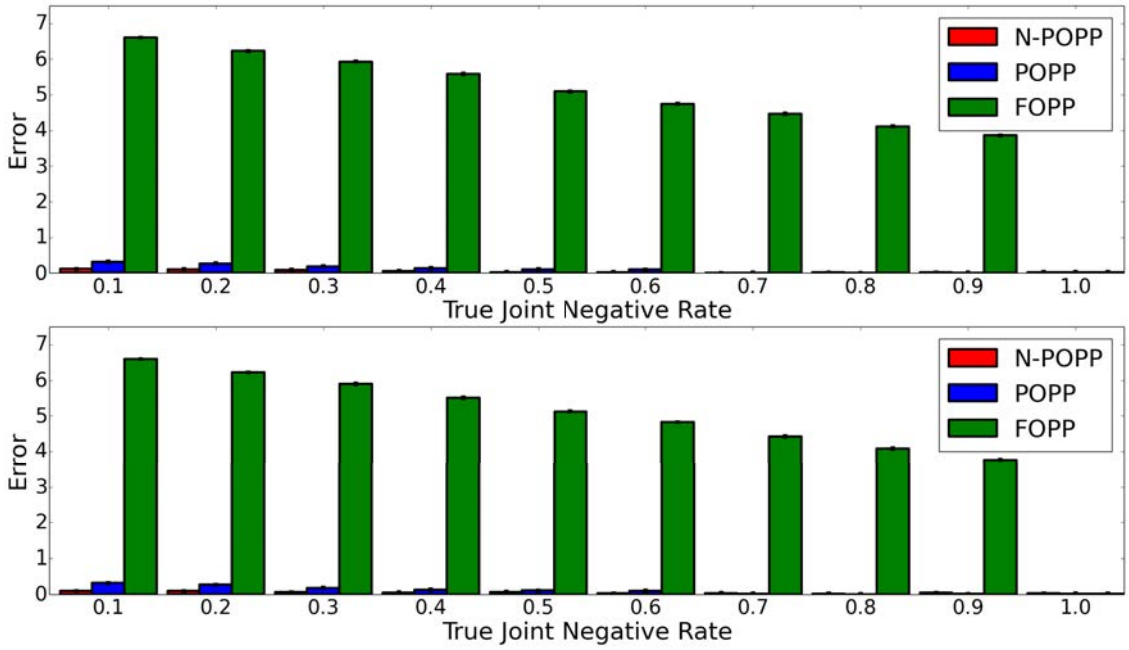

Figure 8.5: The RMSE of posterior estimates of λ for the N-POPP and the POPP model with variation on $\mathscr{E}^-$ with negative correlation between two sensors. Accuracies of MAP estimates in the top panel, accuracies of expectation of the posterior in the bottom panel. Each data point is an average of 30 trials. Standard errors are shown.

specificity in the POPP model) while fixing the true joint positive rate $\mathscr{E}^+$. Once again, the variation falls into three categories:

1. The variation on the true joint negative rate $\mathscr{E}^-$ where two sensors are positively correlated in negative detections. This is shown by $P_{jnt}(d_1 = 0, d_2 = 0 \mid e = 0) > 0.0$, and $P_{jnt}(d_1 = 0, d_2 = 1 \mid e = 0) = P_{jnt}(d_1 = 1, d_2 = 0 \mid e = 0) = 0.0$. The performance of this variation is shown in Figure 8.4.

2. The variation on the true joint negative rate $\mathscr{E}^-$ where two sensors are negatively correlated in false negative detections. This is shown by $P_{jnt}(d_1 = 0, d_2 = 1 \mid e = 0) > 0.0$, $P_{jnt}(d_1 = 1, d_2 = 0 \mid e = 0) > 0.0$, and $P_{jnt}(d_1 = 0, d_2 = 0 \mid e = 0) = 0.0$. The performance of this variation is shown in Figure 8.5.

3. The variation on the true joint negative rate $\mathscr{E}^-$ where two sensors are uncorrelated in negative detections (similar to the POPP model with an independent assumption). This is shown by $P_{jnt}(d_1 = 0, d_2 = 1 \mid e = 0) > 0.0$, $P_{jnt}(d_1 = 1, d_2 = 0 \mid e = 0) > 0.0$, and

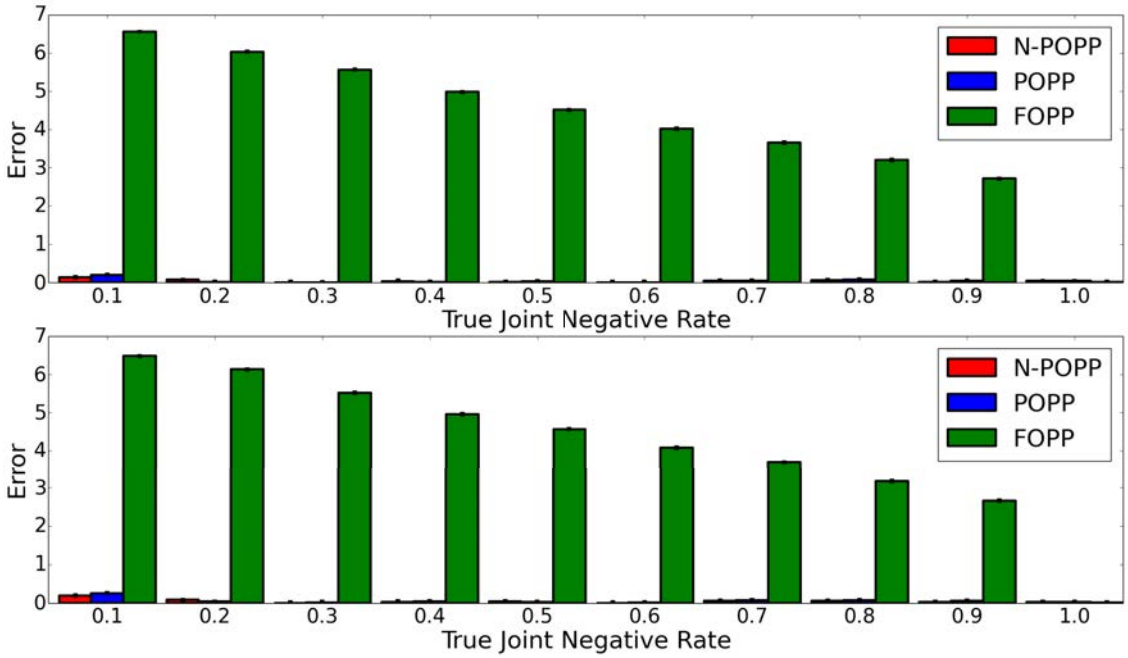

Figure 8.6: The RMSE of posterior estimates of λ for the N-POPP and the POPP model with variation on $\mathscr{E}^-$ with no correlation between two sensors. Each trial consisted of Accuracies of MAP estimates in the top panel, accuracies of expectation of the posterior in the bottom panel. Each data point is an average of 30 trials. Standard errors are shown.

$P_{jnt}(d_1 = 0, d_2 = 0 \mid e = 0) > 0.0$. The performance of this variation is shown in Figure 8.6.

One should note that as it involves more than one sensor, the FOPP filter displayed in all figures above is the result of averaging the MAP hypothesis or the expectation of the posterior over λ from each sensor.

The performance of the N-POPP, the POPP and the FOPP filters were also assessed by comparing the Jensen-Shannon distance of each approximated posterior distribution over λ to the true λ'. Different correlations between two sensors were tested: positive correlation, negative correlation, and no correlation. For each correlation type, a further variation to different level of sensor unreliability was considered. First, two variations were made to the true joint positive rate $\mathscr{E}^+$ (TJPR), while fixing the true joint negative rate $\mathscr{E}^-$ (TJNR) on each type of correlation. This includes:

- $P_{jnt}(d_{1k} = 1, d_{2k} = 1 \mid e_k = 1) = 0.1, P_{jnt}(d_{1k} = 0, d_{2k} = 0 \mid e_k = 1) = 0.9$ ($\mathscr{E}^+$ with low positive correlation);

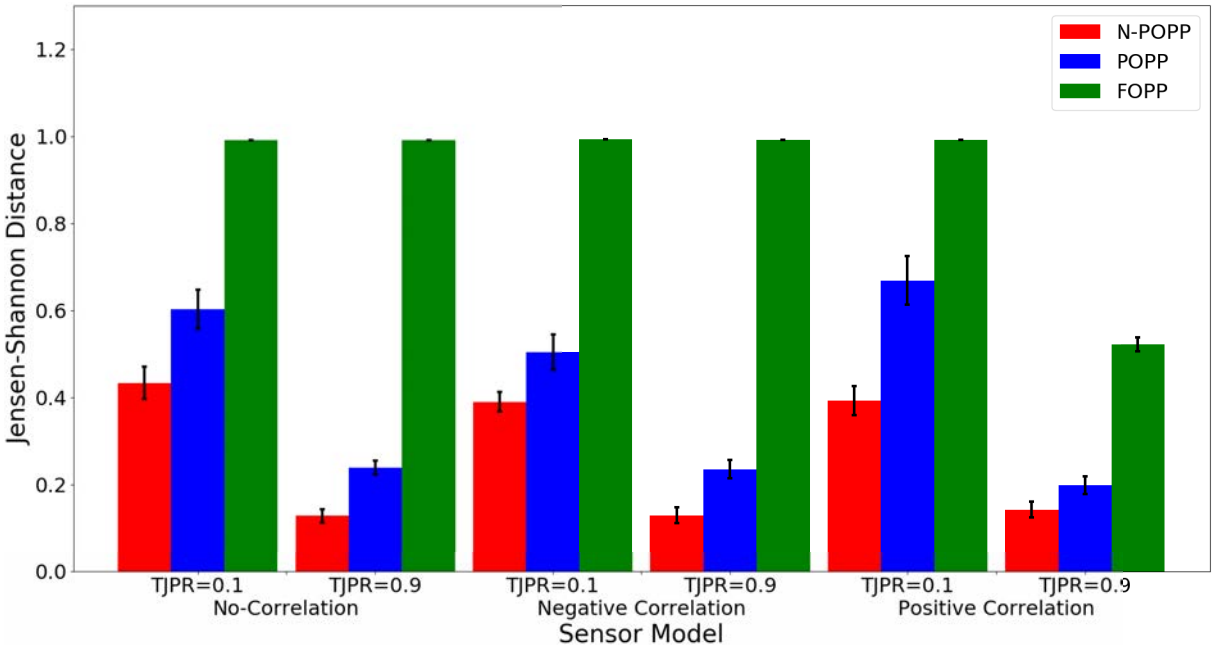

Figure 8.7: The Jensen-Shannon distance of posterior estimates of λ for the N-POPP, the POPP and the FOPP models with variations on the true joint positive rate. Each trial consisted of a stream of $\overrightarrow{s_1} \dots \overrightarrow{s_{144}}$ samples to update $P_G(\lambda \mid \overrightarrow{s_i})$. Standard errors are shown.

- $P_{jnt}(d_{1k} = 1, d_{2k} = 1 \mid e_k = 1) = 0.9, P_{jnt}(d_{1k} = 0, d_{2k} = 0 \mid e_k = 1) = 0.1$ ($\mathscr{E}^+$ with high positive correlation);

- $P_{jnt}(d_{1k} = 1, d_{2k} = 0 \mid e_k = 1) = 0.05, P_{jnt}(d_{1k} = 0, d_{2k} = 1 \mid e_k = 1) = 0.05, P_{jnt}(d_{1k} = 0, d_{2k} = 0 \mid e_k = 1) = 0.9$ ($\mathscr{E}^+$ with low negative correlation);

- $P_{jnt}(d_{1k} = 1, d_{2k} = 0 \mid e_k = 1) = 0.45, P_{jnt}(d_{1k} = 0, d_{2k} = 1 \mid e_k = 1) = 0.45, P_{jnt}(d_{1k} = 0, d_{2k} = 0 \mid e_k = 1) = 0.1$ ($\mathscr{E}^+$ with high negative correlation);

- $P_{jnt}(d_{1k} = 1, d_{2k} = 0 \mid e_k = 1) = 0.033, P_{jnt}(d_{1k} = 0, d_{2k} = 1 \mid e_k = 1) = 0.033, P_{jnt}(d_{1k} = 1, d_{2k} = 1 \mid e_k = 1) = 0.033, P_{jnt}(d_{1k} = 0, d_{2k} = 0 \mid e_k = 1) = 0.901$ ($\mathscr{E}^+$ with no correlation – Similar to a sensor model with low TPR);

- $P_{jnt}(d_{1k} = 1, d_{2k} = 0 \mid e_k = 1) = 0.3, P_{jnt}(d_{1k} = 0, d_{2k} = 1 \mid e_k = 1) = 0.3, P_{jnt}(d_{1k} = 1, d_{2k} = 1 \mid e_k = 1) = 0.3, P_{jnt}(d_{1k} = 0, d_{2k} = 0 \mid e_k = 1) = 0.1$ ($\mathscr{E}^+$ with no correlation – Similar to a sensor model with moderate TPR).

The assessment of the filters in terms of the Jensen-Shannon distance with the above testings

can be seen in Figure 8.7.

Second, two variations to the true joint negative rate $\mathscr{E}^-$ (TJNR), while fixing the true joint positive rate $\mathscr{E}^+$ (TJPR) on each type of correlation. This includes:

- $P_{jnt}(d_{1k}=1,d_{2k}=1 \mid e_k=0)=0.1, P_{jnt}(d_{1k}=0,d_{2k}=0 \mid e_k=0)=0.9$ ($\mathscr{E}^-$ with low positive correlation);
- $P_{jnt}(d_{1k}=1,d_{2k}=1 \mid e_k=0)=0.9, P_{jnt}(d_{1k}=0,d_{2k}=0 \mid e_k=0)=0.1$ ($\mathscr{E}^-$ with high positive correlation);
- $P_{jnt}(d_{1k}=1,d_{2k}=0 \mid e_k=0)=0.05, P_{jnt}(d_{1k}=0,d_{2k}=1 \mid e_k=0)=0.05, P_{jnt}(d_{1k}=0,d_{2k}=0 \mid e_k=0)=0.9$ ($\mathscr{E}^-$ with low negative correlation);
- $P_{jnt}(d_{1k}=1,d_{2k}=0 \mid e_k=0)=0.45, P_{jnt}(d_{1k}=0,d_{2k}=1 \mid e_k=0)=0.45, P_{jnt}(d_{1k}=0,d_{2k}=0 \mid e_k=0)=0.1$ ($\mathscr{E}^-$ with high negative correlation);
- $P_{jnt}(d_{1k}=1,d_{2k}=0 \mid e_k=0)=0.033, P_{jnt}(d_{1k}=0,d_{2k}=1 \mid e_k=0)=0.033, P_{jnt}(d_{1k}=1,d_{2k}=1 \mid e_k=0)=0.033, P_{jnt}(d_{1k}=0,d_{2k}=0 \mid e_k=1)=0.901$ ($\mathscr{E}^-$ with no correlation – Similar to a sensor model with low TNR);
- $P_{jnt}(d_{1k}=1,d_{2k}=0 \mid e_k=0)=0.3, P_{jnt}(d_{1k}=0,d_{2k}=1 \mid e_k=0)=0.3, P_{jnt}(d_{1k}=1,d_{2k}=1 \mid e_k=0)=0.3, P_{jnt}(d_{1k}=0,d_{2k}=0 \mid e_k=1)=0.1$ ($\mathscr{E}^-$ with no correlation – Similar to a sensor model with moderate TNR).

The assessment of the filters in terms of the Jensen-Shannon distance with the above testings can be seen in Figure 8.8.

Having done a thorough evaluation on the performance of the N-POPP filter against the standard POPP filter, it is safe to conclude that the N-POPP filter clearly outperforms the POPP filter in all given scenarios in estimating λ parameter of a Poisson process. It outperforms the FOPP estimator by a big margin. Since these scenarios are similar to the scenarios applied in the standard POPP model, it is safely to assume that adding more sensors to the experiment does not change the performance ratio between the N-POPP model and the POPP model.

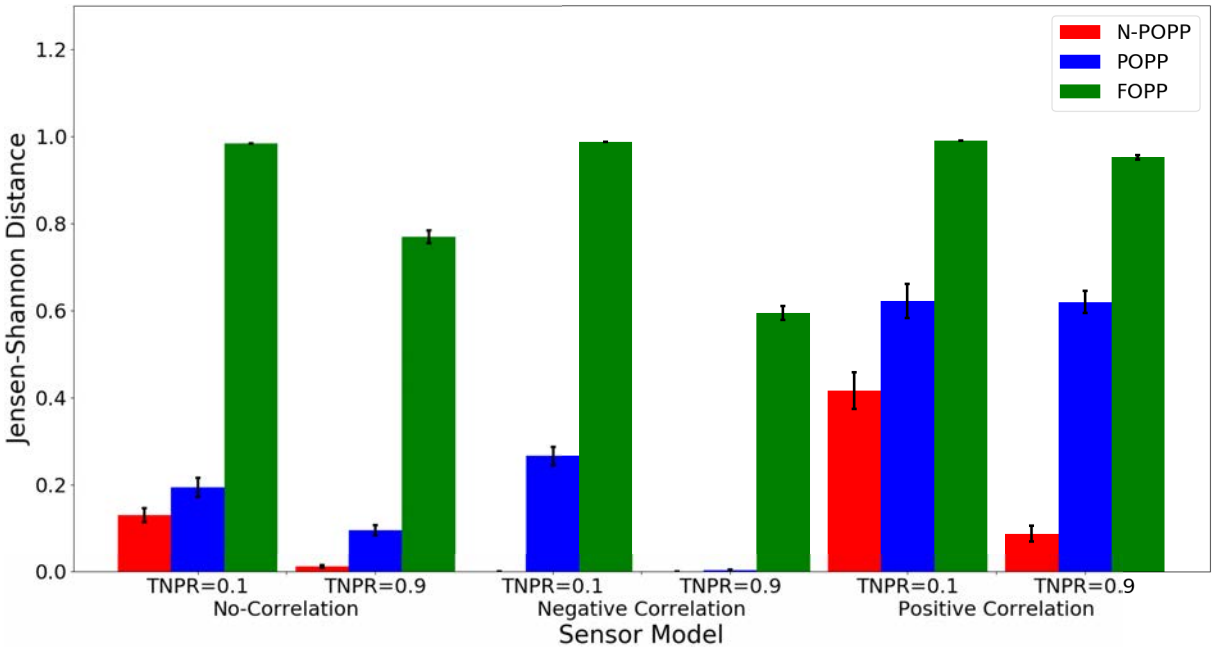

Figure 8.8: The Jensen-Shannon distance of posterior estimates of λ for the N-POPP, the POPP and the FOPP models with variations on the true joint negative rate. Each trial consisted of a stream of $\overrightarrow{s_1} \ldots \overrightarrow{s_{144}}$ samples to update $P_G(\lambda \mid \overrightarrow{s_i})$. Standard errors are shown.

8.3 Evaluation on A Real World Dataset

In this evaluation, the N-POPP filter is compared to the POPP-Beta filter, the standard POPP filter, and the FOPP filter in estimating the λ parameter of both a homogeneous Poisson process and a periodic Poisson process on a real world dataset. Similar to the evaluation on the simulated data set, the switching filter is used for all the models except the standard FOPP model.

The extended dataset is used which was explained and used in the previous chapter. Three perception algorithms for a mobile robot counting the number of people passing by in an office building are the leg detector (LD), upper body detector (UBD), and change detector (CD). Each minute, each detector returns a binary value whether there was a person or not. All binary detections coming from the same minute are grouped together, and these groups are further combined into a 10-minute joint detection to form a joint sensed count of the number of people all sensors detect in each 10 minute interval during the day. The detections are organised according to time/date and the spatial region where each detection was made.

Table 8.1: Averaged joint sensor model across all areas trained from 48 days of data.

	$P(d_{ld}, d_{ubd}, d_{cd} \mid e = 1)$	$P(d_{ld}, d_{ubd}, d_{cd} \mid e = 0)$
$d_{ld} = 0, d_{ubd} = 0, d_{cd} = 0$	0.029	0.823
$d_{ld} = 0, d_{ubd} = 0, d_{cd} = 1$	0.373	0.064
$d_{ld} = 0, d_{ubd} = 1, d_{cd} = 0$	0.016	0.055
$d_{ld} = 0, d_{ubd} = 1, d_{cd} = 1$	0.092	0.008
$d_{ld} = 1, d_{ubd} = 0, d_{cd} = 0$	0.140	0.013
$d_{ld} = 1, d_{ubd} = 0, d_{cd} = 1$	0.081	0.001
$d_{ld} = 1, d_{ubd} = 1, d_{cd} = 0$	0.104	0.033
$d_{ld} = 1, d_{ubd} = 1, d_{cd} = 1$	0.165	0.003

The sensor models for both the POPP filter and the POPP-Beta filter, and the joint sensor model for the N-POPP filter were built from sensor counts and true counts. To build the sensor models and measure the performance of each filter including the FOPP filter, four fold cross-validation was performed on the 48-day dataset with the unit being the whole days, i.e., 12 days of data was used to build the sensor models and the remaining 36 days of data was used as a test set. Similarly to the previous chapter, this process is done for all areas within the patrol space. An example of a joint sensor model for the N-POPP filter trained from 48 days of data can be seen in Table 8.1.

An estimated true λ for the Poisson process and $\lambda(t_i, t_j)$ for the periodic Poisson process on each region is approximated by running a FOPP filter on the true counts, whereas the uncorrected estimate λ and $\lambda(t_i, t_j)$ according to the FOPP model was estimated only from the change detector count data. Using RMSE and the Jensen-Shannon distance as distance metrics, the estimated true λ is then used as a comparison to the resulting estimated λ produced by each filter. One should note that the FOPP filter estimated the λ based only on the change detector count data since it has been shown in Table 6.3 that the change detector has better accuracy. The comparison was done for all areas within the patrol space.

8.3.1 The N-POPP models on Homogeneous Poisson Processes

This section starts with a case on homogeneous Poisson processes with an interval of 10 minute. The results are shown in Figure 8.9 for the RMSE and 8.10 for the Jensen-Shannon

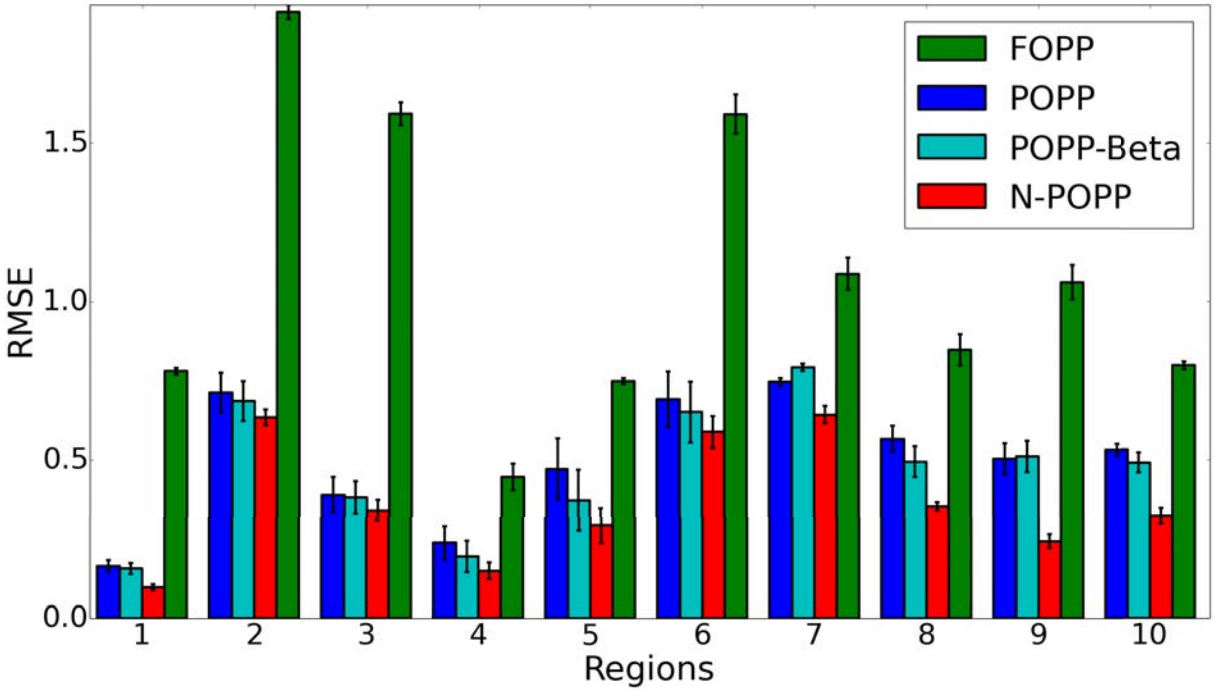

Figure 8.9: The RMSE of the FOPP, POPP, POPP-Beta, N-POPP estimators of λ as it varies across areas (regions) of the environment. Standard error is shown.

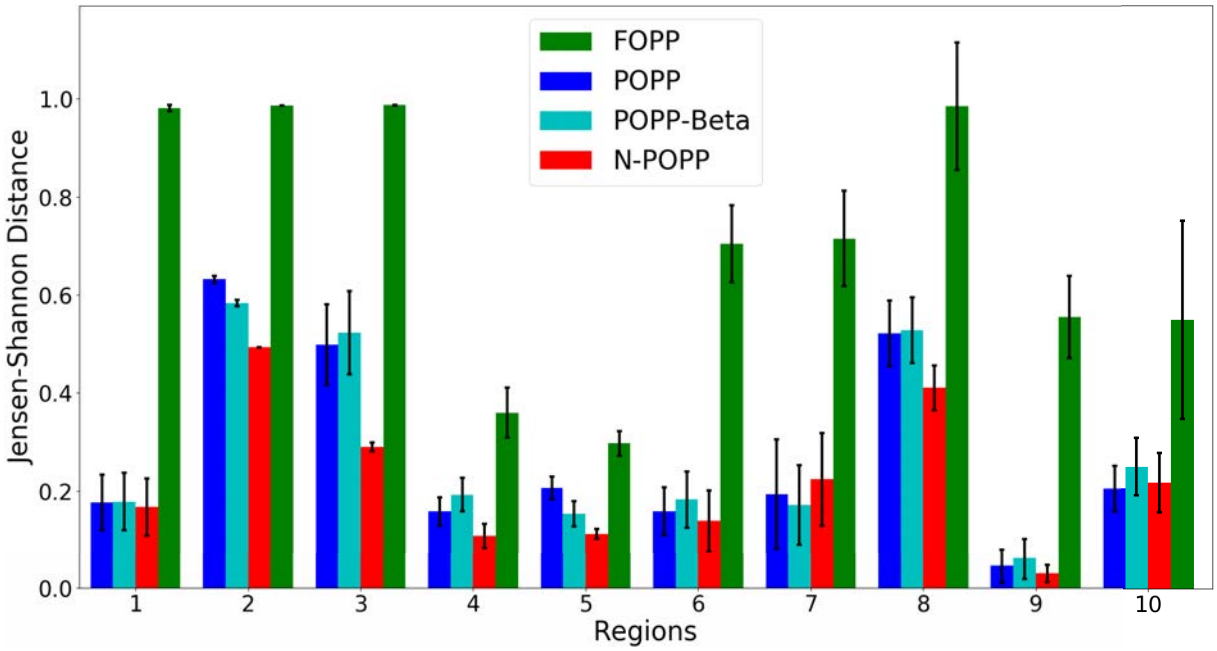

Figure 8.10: The Jensen-Shannon distance of the FOPP, the POPP, the POPP-Beta, and the N-POPP model distributions of λ as it varies across areas (regions) of the environment. Standard error is shown.

distance. These figures are an extension to Figure 7.5 and 7.6 by adding an evaluation of the N-POPP filter. As can be seen from both figures, the N-POPP model produces both better

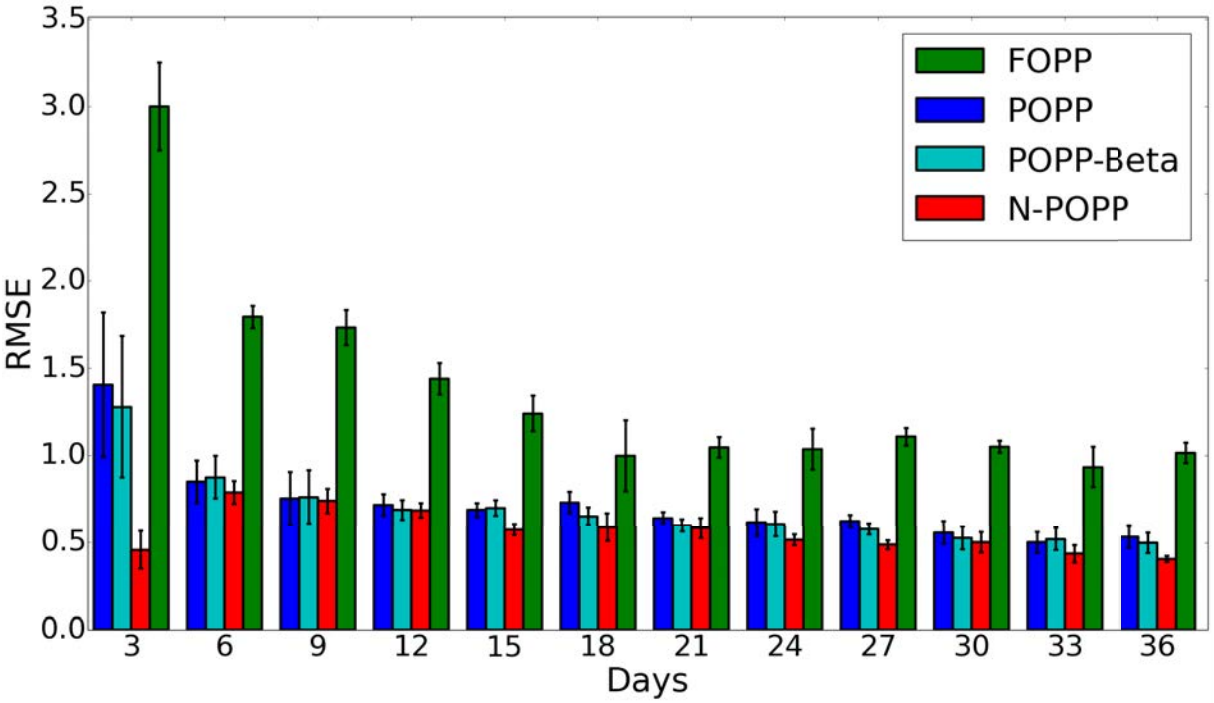

Figure 8.11: The RMSE evolution from day 3 to day 36 with 3 day interval, averaged across all regions. Standard error is shown.

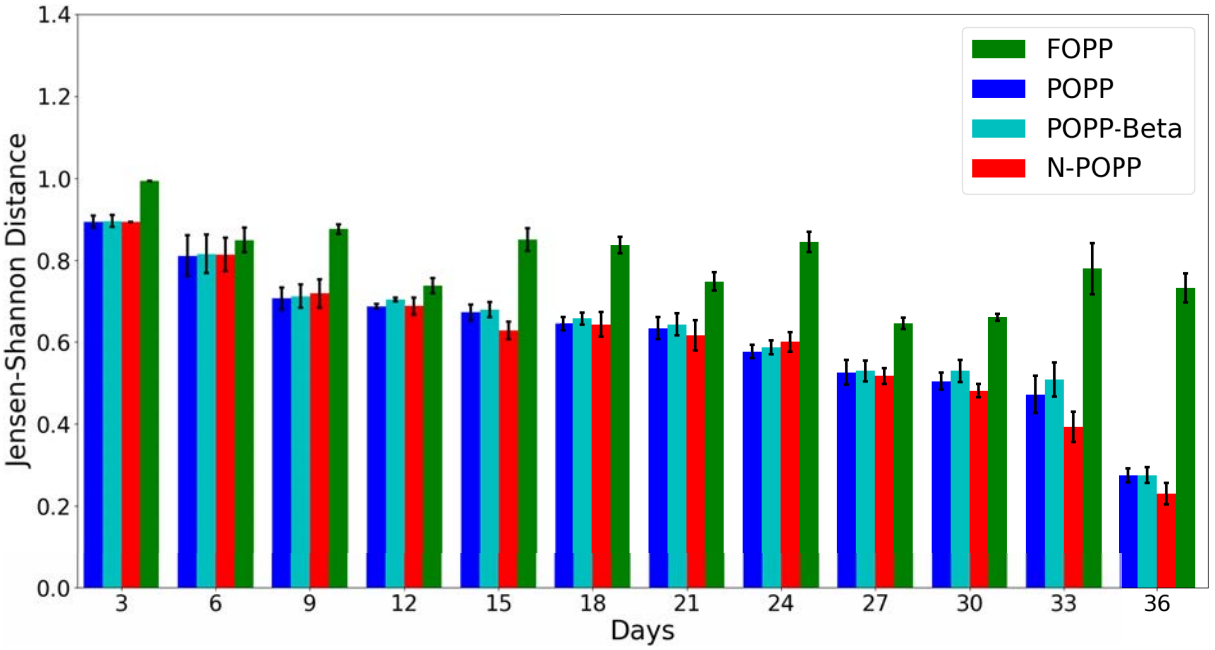

Figure 8.12: The Jensen-Shannon distance evolution of the FOPP, the POPP, the POPP-Beta, and the N-POPP model distributions of λ from day 3 to day 36 with 3 day interval, averaged across all regions. Standard error is shown.

estimates and more similar distributions than both the POPP-Beta model, and the POPP model. Recall that the distribution of detections is not uniform across areas, some areas

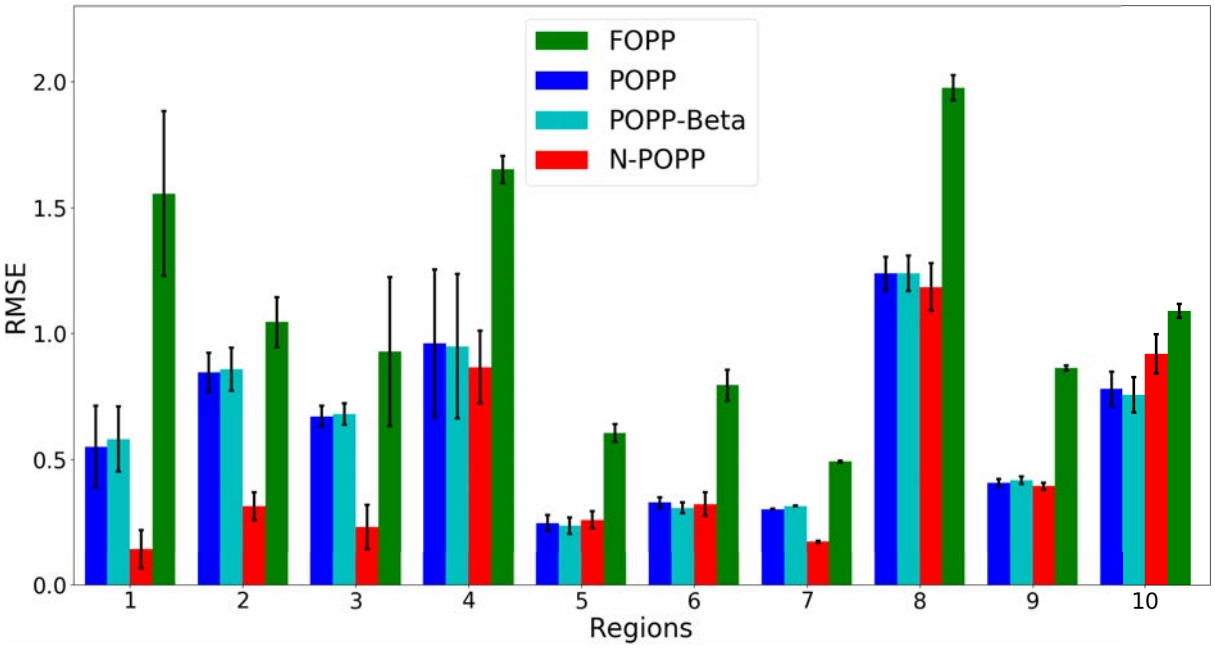

Figure 8.13: The RMSE of the FOPP, POPP, POPP-Beta, and N-POPP estimators of $\lambda(t_i, t_j)$ as it varies across areas (regions) of the environment. Standard error is shown.

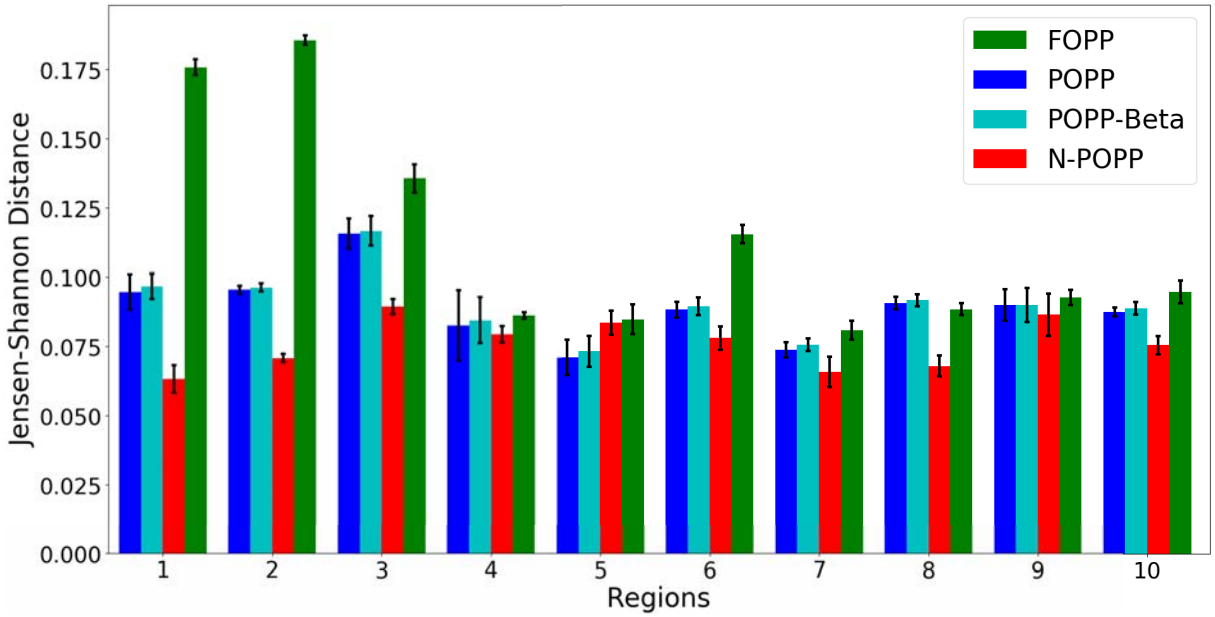

Figure 8.14: The Jensen-Shannon distance of the FOPP, POPP, POPP-Beta, and N-POPP model distributions of $\lambda(t_i, t_j)$ as it varies across areas (regions) of the environment. Standard error is shown.

(like area 7 and 9) have limited data. This, however, does not affect the N-POPP filter from producing a better estimate than the other filters.

An evaluation on how the N-POPP evolved with time was also conducted, again in terms

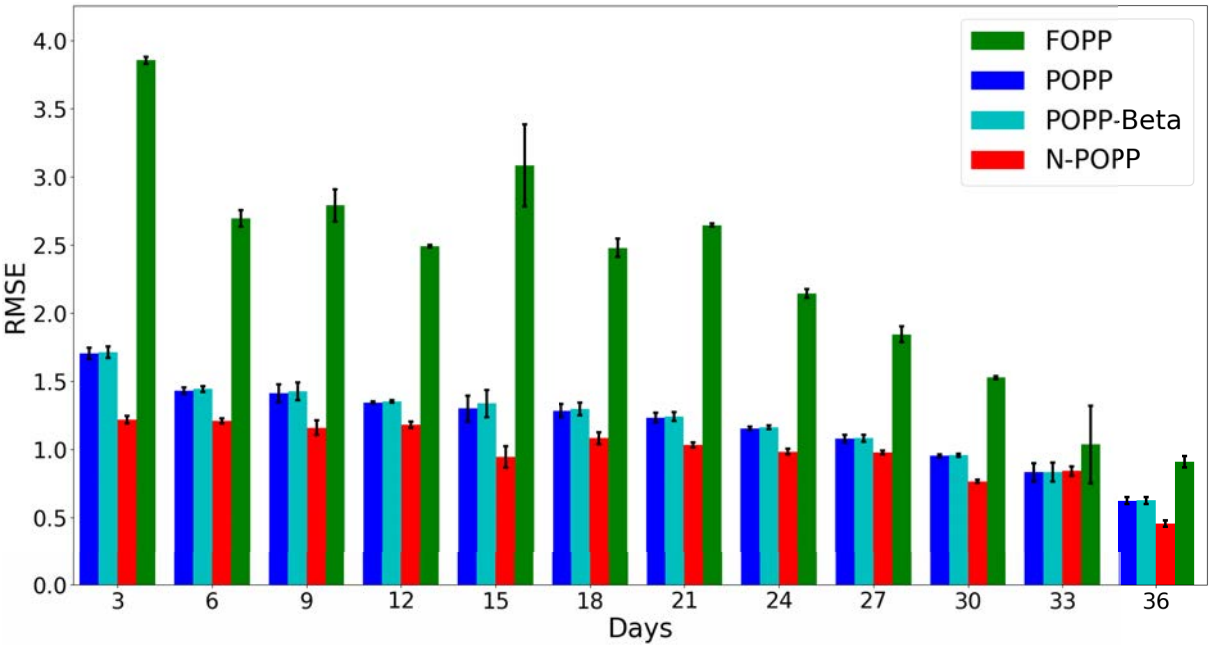

Figure 8.15: The RMSE evolution of periodic Poisson processes with POPP, POPP-Beta, N-POPP and FOPP filters from day 3 to day 36, averaged across all regions. Standard error is shown.

of both RMSE and Jensen-Shannon distance, since it is expected to gradually get closer to the true λ. Figures 8.11 and 8.12 show that as time passes the performance of the N-POPP estimator steadily becomes better. Once again, the figures are extensions of Figures 7.7 and 7.8.

8.3.2 The N-POPP models on Periodic Poisson Processes

This evaluation, a periodic Poisson process in which the imposed single periodicity is a one-day cycle, i.e. $\lambda(t_i, t_j) = \lambda(t_{i+\Delta}, t_{j+\Delta})$ with $\Delta = 24 * 60$(minutes), is assumed to be the underlying process.

Figures 8.13 and 8.14 showcase that the N-POPP model is able to cope and overcome the problems with limited sample data both for building the (joint) sensor model and estimating the $\lambda(t_i, t_j)$. This is showcased in regions 4, 7, and 9 where the N-POPP filter were more accurate than the POPP and the POPP-Beta filter. In many other regions, the N-POPP models still managed to display better estimates than the POPP and the POPP-Beta filters. This infers

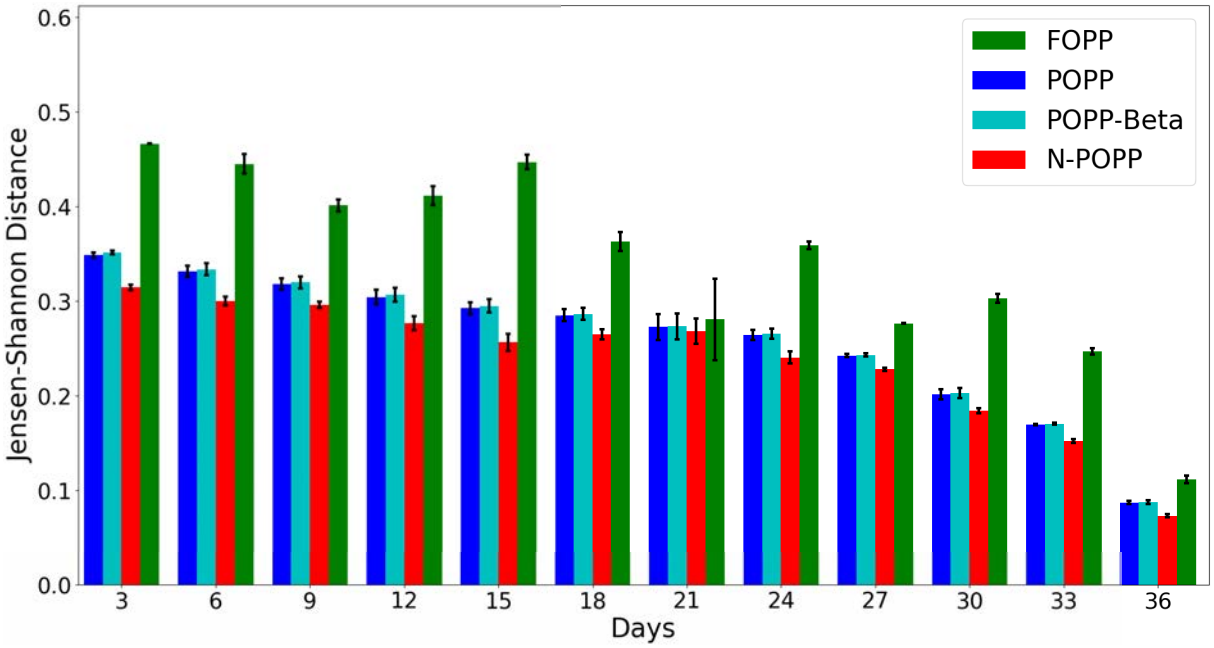

Figure 8.16: The Jensen-Shannon distance evolution of the FOPP, the POPP, the POPP-Beta, and the N-POPP filters in periodic Poisson processes from day 3 to day 36 in a 3-day interval, averaged across all regions. Standard error is shown.

that the N-POPP performed much better than the standard Bayesian inference according to the FOPP model.

Once again, a side evaluation on how the N-POPP evolved with time was done. Figures 8.15 and 8.16 show the evolution of each filter overtime in terms of RMSE and Jensen-Shannon distance across regions. The N-POPP filter gradually increased its accuracy in estimating $\lambda(t_i, t_j)$. The filter constantly outperformed the POPP and the POPP-Beta filters in terms of accuracy. This transitively implies that the N-POPP filter is a better estimator than the rest of the filters.

8.4 Discussion

In summary, this chapter introduced the second extension of the partially observable Poisson process, which removes the assumption that the sensors are uncorrelated to one another. In the first part of this chapter, a mathematical model of the non-independent POPP (N-POPP) was defined. The proposed model extends the sensor model used by the POPP and the

POPP-Beta models into a joint sensor model where the outcomes are the combined detection patterns across the sensors.

In the second part of the chapter, an evaluation to the N-POPP model was carried out. The experiments on simulated data showed that the extension of the POPP model–the N-POPP–was more accurate in estimating the parameter λ of a Poisson process than the POPP model. A further evaluation on real-world data was done. This evaluation included the extension detailed in the previous chapter, the POPP-Beta model. The result showed that the N-POPP model outperformed the other models including both the POPP and the POPP-Beta model. Not only that, the N-POPP model managed to outperform the FOPP model, where the POPP and the POPP-Beta have failed, in the periodic Poisson process cases where sample data were limited.

The POPP model has been extended to relax the assumptions of the POPP model. Up to this chapter, both two underlying assumptions are tackled separately where one assumption is tackled by the POPP-Beta model, and the other is tackled by the N-POPP model. This leaves us with how to combine the benefits of each correction.

CHAPTER 9

POPP: NON-INDEPENDENT SENSOR FAILURES WITH UNRELIABLE SENSOR MODELS

In the last two chapters, two limitations of the standard POPP model were studied. The problem of requiring a precise sensor model was studied and evaluated in Chapter 7. The problem of having an assumption that counter failures are conditionally independent from one to another given the true count was studied and evaluated in the last chapter. Each of these problems has been studied and evaluated separately.

This chapter focuses on merging the solutions provided by the last two chapters into one. The result of this merging is a new model which is an extension of the POPP model without either limitation that the POPP model has. The new model is based on the N-POPP model introduced in Chapter 8 where the joint sensor model is modelled as latent variables, hence, updating Equation 8.6.

The N-POPP definition in Section 9.1 is revisited. Section 9.2 introduces an extension to the N-POPP model where the degree of confidence of a joint sensor model is modelled

and a prior knowledge over the sensor's confidence is incorporated. The performance of this extension is tested on simulated data, and the result is explained in Section 9.3. Section 9.4 uses the real world data from the previous chapter and extends the evaluation of the proposed model on the real world data, and shows the results of the POPP extension on the real world data. Section 9.5 closes this chapter by concluding the contribution.

9.1 Non-Independent Partially Observable Poisson Process (N-POPP)

The N-POPP model is a POPP model where counts from multiple sensors are jointly corrected. Similar to the POPP model, there is still a separation between the true count x_i and sensed count $\overrightarrow{s_i}$ in every time interval i. However, the sensed count $\overrightarrow{s_i}$ has been redefined to:

$$\overrightarrow{s_i} = (D_0 = g_0, \ldots, D_{(m^2)-1} = g_{(m^2)-1}) \tag{9.1}$$

with m as the number of sensors, D_0 as a representation of $(d_{1k} = 0, \ldots, d_{mk} = 0)$, D_1 as a representation of $(d_{1k} = 0, \ldots, d_{mk} = 1)$, …, $D_{(m^2)-1}$ as a representation of $(d_{1k} = 1, \ldots, d_{mk} = 1)$, $g_0, \ldots, g_{(m^2)-1} \in \mathbb{N}$, $d_{jk} \in [0,1]$, and $1 \leq j \leq m$.

The main difference between the POPP model and the N-POPP model relies on the definition of the sensed count likelihood $P(\overrightarrow{s_i} \mid x_i)$. The probability of $\overrightarrow{s_i}$ set of detections given x_i events happened has become the aggregate of joint detections given the positive event $P_{jnt}(d_{1k}, \ldots, d_{mk} \mid e_k = 1)$ in x_i sub-intervals, and joint detections given the negative event $P_{jnt}(d_{1k}, \ldots, d_{mk} \mid e_k = 0)$ in $l - x_i$ sub-intervals. Formally, it is defined as:

$$P(\overrightarrow{s_i} \mid x_i) = \sum_{\overrightarrow{s^+} \subseteq \overrightarrow{s_i}} Multi(\overrightarrow{s^+} \mid x_i, \mathscr{E}^+)\ Multi(\overrightarrow{s^-} \mid (l - x_i), \mathscr{E}^-) \tag{9.2}$$

with l as the number of subintervals between (t_1, t_2), $\| \overrightarrow{s^+} \| = x_i$, $\| \overrightarrow{s_i} \| = \| \overrightarrow{s^+} \| + \| \overrightarrow{s^-} \|$, $\mathscr{E}^+ = (E_0^+, \ldots, E_{(m^2)-1}^+)$, $\mathscr{E}^- = (E_0^-, \ldots, E_{(m^2)-1}^-)$, $E_0^+ = P(d_{1k} = 0, \ldots, d_{mk} = 0 \mid e_k = 1), \ldots, E_{(m^2)-1}^+ = P(d_{1k} = 1, \ldots, d_{mk} = 1 \mid e_k = 1)$, and $E_0^- = P(d_{1k} = 0, \ldots, d_{mk} = 0 \mid e_k = 0), \ldots, E_{(m^2)-1}^- = P(d_{1k} =$

$1, \ldots, d_{mk} = 1 \mid e_k = 0)$.

Apart from this equation, the N-POPP model follows the standard POPP model. Consequently, the posterior of λ, given n samples $\overrightarrow{s} = (\overrightarrow{s_1} \ldots \overrightarrow{s_n})$ still follows the definition in Equation 6.2. A graphical model of the N-POPP also follows the POPP model and is shown in Figure 6.1.

9.2 N-POPP with Unreliable Joint Sensor Model (POPP-Dirichlet)

As explained in the previous section, the N-POPP model uses a joint sensor model in estimating the parameter λ of a Poisson process. A joint sensor model is an extension of a sensor model where the model provides a probability for a particular combination of binary detections coming from each sensor given the true event. Unlike a standard sensor model for sensor j which is represented by its true positive rate $tpr_j = P_j(d_k = 1 \mid e_k = 1)$ and its false positive rate $fpr_j = P_j(d_k = 1 \mid e_k = 0)$, a joint sensor model is represented as a joint detection among sensors given the true event as:

$$P_{jnt}(d_{1k}, \ldots, d_{mk} \mid e_k) \tag{9.3}$$

with d_{jk} being the detection of sensor j at subinterval I_k, e_k being the true event, and $d_{jk}, e_k \in [0,1]$.

To construct a joint sensor model, one needs to have both detections and the corresponding actual (non)-event as ground truth. Pre-processing involving expert interventions is typically required before the detections and their corresponding ground truth can be further used. Similar to the POPP model, the N-POPP model requires the joint sensor model to be accurate to avoid the posterior over λ to drift away from the true posterior. If attaining an accurate sensor model for the POPP model is a problem, then this becomes more prominent in the case of a joint sensor model. This is because the training data needed to construct a joint

sensor model grows by a factor of two for each sensor involved. Similar to the transformation from the POPP model to the POPP-Beta model, the joint sensor model is transformed into a Bayesian estimation problem, where $P_{jnt}(d_{1k},\ldots,d_{mk} \mid e_k = 1)$ and $P_{jnt}(d_{1k},\ldots,d_{mk} \mid e_k = 0)$ of a joint sensor model follow Dirichlet distributions. The Dirichlet distribution is chosen due to its computational convinience as Dirichlet distributions provide a family of prior probability distributions for the multinomial distributions where $P_{jnt}(d_{1k},\ldots,d_{mk} \mid e_k = 1)$ and $P_{jnt}(d_{1k},\ldots,d_{mk} \mid e_k = 0)$ set the probabilities of multinomial distributions in Equation 9.2. The dirichlet-multinomial conjugacy leads to an analytically tractable compound distribution which is called the Dirichlet-multinomial distribution, where the $\overrightarrow{p} = (p_1,\ldots,p_r)$ parameter in the multinomial distribution $Multi((d_1,\ldots,d_r) \mid x, \overrightarrow{p})$ is randomly drawn from a Dirichlet distribution $Dir(\overrightarrow{p} \mid (\zeta_1,\ldots,\zeta_r))$.

$$
\begin{aligned}
P((d_1,\ldots,d_r) \mid x,(\zeta_1,\ldots,\zeta_r)) &= \int P((d_1,\ldots,d_r) \mid x,\overrightarrow{p})\; P(\overrightarrow{p} \mid (\zeta_1,\ldots,\zeta_r))\; d\mathbb{S}_r \\
&= \int Multi((d_1,\ldots,d_r) \mid x,\overrightarrow{p})\; Dir(\overrightarrow{p} \mid (\zeta_1,\ldots,\zeta_r))\; d\mathbb{S}_r \\
&= \int \frac{\Gamma(x+1)}{\prod_{j=1}^{r}\Gamma(d_j+1)} \prod_{j=1}^{r} p_j^{d_j} \frac{\Gamma(\sum_{j=1}^{r}\zeta_j)}{\prod_{j=1}^{r}\Gamma(\zeta_j)} \prod_{j=1}^{r} p_j^{\zeta_j-1}\, d\mathbb{S}_r \\
&= \int \frac{\Gamma(x+1)}{\prod_{j=1}^{r}\Gamma(d_j+1)} \frac{\Gamma(\sum_{j=1}^{r}(\zeta_j))}{\prod_{j=1}^{r}\Gamma(\zeta_j)} \int \prod_{j=1}^{r} p_j^{(\zeta_j+d_j)-1}\, d\mathbb{S}_r \\
&= \frac{\Gamma(x+1)}{\prod_{j=1}^{r}\Gamma(d_j+1)} \frac{\Gamma(\sum_{j=1}^{r}(\zeta_j))}{\prod_{j=1}^{r}\Gamma(\zeta_j)} \frac{\prod_{j=1}^{r}\Gamma(d_j+\zeta_j)}{\Gamma(x+\sum_{j=1}^{r}\zeta_j)} \\
&= DM((d_1,\ldots,d_r) \mid x,(\zeta_1,\ldots,\zeta_r))
\end{aligned}
\tag{9.4}
$$

with $d\mathbb{S}_r$ denotes integrating $\overrightarrow{p}$ with respect to the $(r-1)$ simplex[1]. One should note that the Dirichlet distribution can be replaced with any other distribution. However, a different approach which involves complicated integral on $P(x,(\zeta_1,\ldots,\zeta_r))$ and a numerical approximation to $P(x,(\zeta_1,\ldots,\zeta_r))$ must be taken into consideration. Avoiding extra computation, either to calculate $P(x,(\zeta_1,\ldots,\zeta_r))$ or to approximate it, and an extra approximation to the posterior

[1]The support of the Dirichlet distribution is the $(r-1)$-dimensional simplex $\mathbb{S}_r$; that is, all r dimensional vectors which form a valid probability distribution

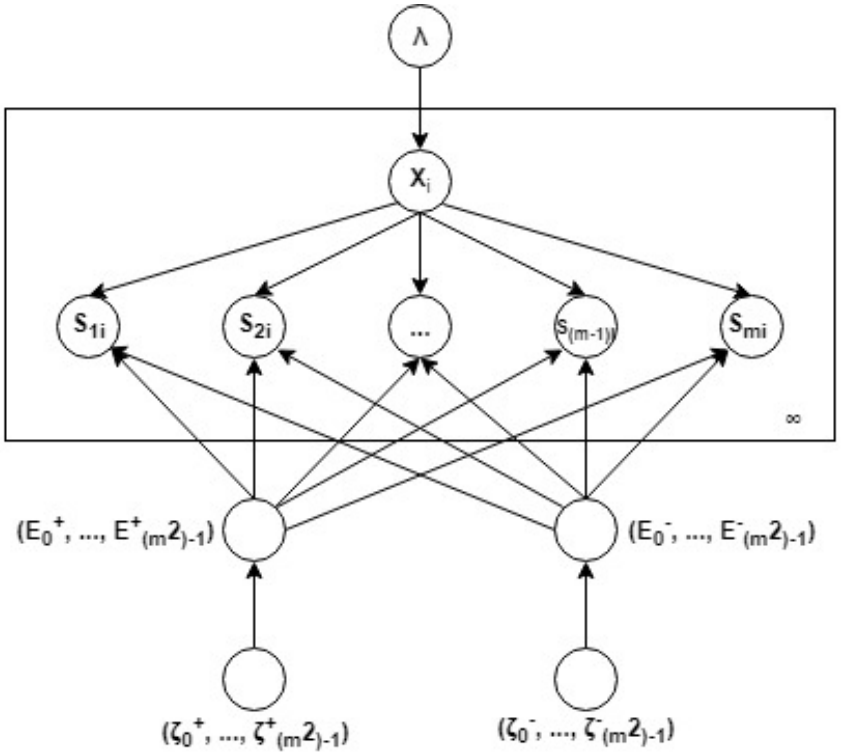

Figure 9.1: Graphical representation of the POPP-Dirichlet.

distribution $P_G(\lambda \mid \overrightarrow{s})$ makes the Dirichlet distribution with its conjugacy a sensible choice.

Given m sensors, a joint sensor model is now represented as two Dirichlet distributions: $Dir(\mathscr{E}^+ \mid \overrightarrow{\zeta^+})$, and $Dir(\mathscr{E}^- \mid \overrightarrow{\zeta^-})$ with $\overrightarrow{\zeta^+} = (\zeta_0^+, \ldots, \zeta_{(m^2)-1}^+)$ and $\overrightarrow{\zeta^-} = (\zeta_0^-, \ldots, \zeta_{(m^2)-1}^-)$. $\overrightarrow{\zeta^+}$ and $\overrightarrow{\zeta^-}$set the overall shape of the Dirichlet priors. $\overrightarrow{\zeta^+}$ can loosely be interpreted as the number of times combination ζ_q^+ of m-sensor detection, $1 \leq q \leq (m^2-1)$, has been made given the positive detection e ($e = 1$). $\overrightarrow{\zeta^-}$ can loosely be interpreted as the number of times combination ζ_q^- of m-sensor detection, $1 \leq q \leq (m^2-1)$, has been made given the negative detection e ($e = 0$).

Given a joint sensor model where the elements of it follow a Dirichlet density and several Dirichlet-multinomial distributions, which provide an unconditional distribution of $(d_1, \ldots, d_r)$, we replace Equation 9.2 with:

$$P(\overrightarrow{s_i} \mid x_i) = \sum_{\overrightarrow{s^+} \subseteq \overrightarrow{s_i}} DM(\overrightarrow{s^+} \mid x_i, \overrightarrow{\zeta^+})\, DM(\overrightarrow{s^-} \mid (l - x_i), \overrightarrow{\zeta^-}) \tag{9.5}$$

with $\| \overrightarrow{s^+} \| = x_i$, $\| \overrightarrow{s_i} \| = \| \overrightarrow{s^+} \| + \| \overrightarrow{s^-} \|$, $\overrightarrow{\zeta^+} = (\zeta_0^+, \ldots, \zeta_{(m^2)-1}^+)$, $\overrightarrow{\zeta^-} = (\zeta_0^-, \ldots, \zeta_{(m^2)-1}^-)$, $\zeta_0^+ = \#(d_{1k} =$

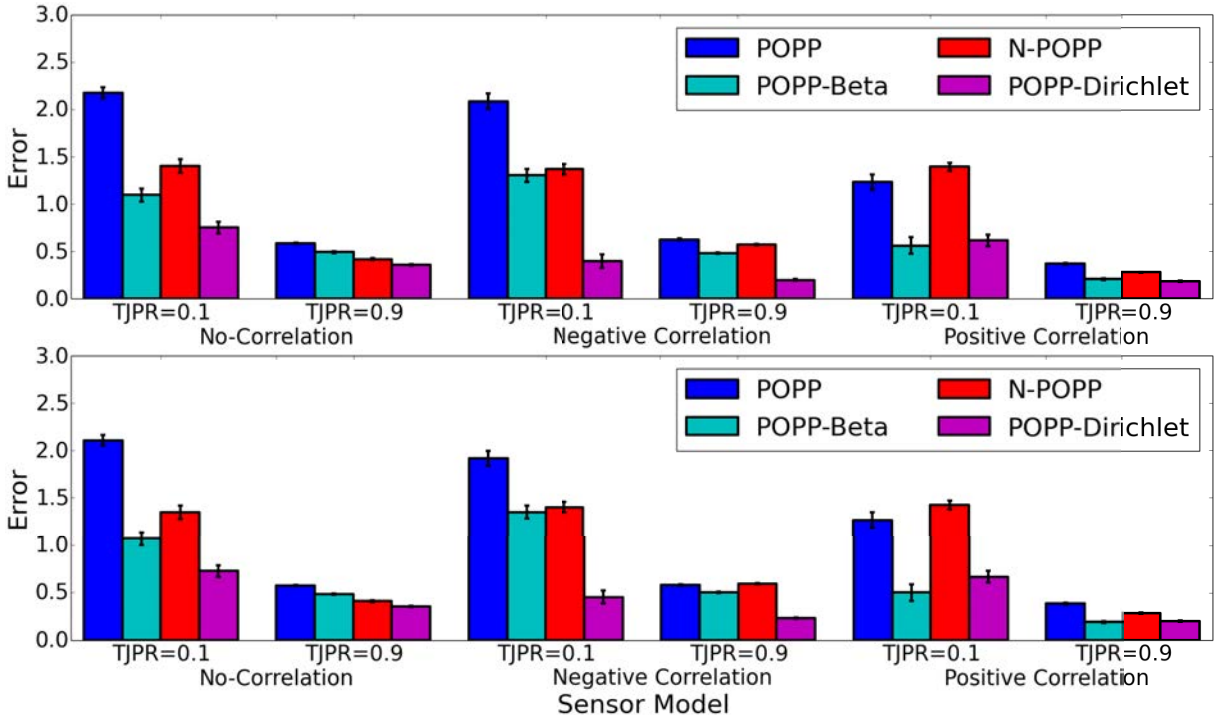

Figure 9.2: The RMSE of posterior estimates of λ for the POPP-Dirichlet and other POPP models with 120 sample data used to build the (joint) sensor model with variation on $\mathscr{E}^+$. Each trial consisted of a stream of $\overrightarrow{s_1} \ldots \overrightarrow{s_{144}}$ samples to update $P_G(\lambda \mid \overrightarrow{s_i})$. Accuracies of MAP estimates are shown in the top panel, accuracies of expectation of the posterior in the bottom panel. Each data point is an average of 30 trials. Standard errors are shown.

$0, \ldots, d_{mk} = 0, e_k = 1), \ldots, \zeta^+_{(m^2)-1} = \#(d_{1k} = 1, \ldots, d_{mk} = 1, e_k = 1)$, and $\zeta^-_0 = \#(d_{1k} = 0, \ldots, d_{mk} = 0, e_k = 0), \ldots, \zeta^-_{(m^2)-1} = \#(d_{1k} = 1, \ldots, d_{mk} = 1, e_k = 0)$.

With a joint sensor model following the Dirichlet density, which is conjugated with multinomial distributions into a posterior predictive distribution shown in Equation 9.5, a graphical model is shown in Figure 9.1. The difference between the N-POPP model and the POPP-Dirichlet lies only in Equations 9.2 and 9.5. However, given a certain Dirichlet prior, and a little training data for the sensor model, the POPP-Dirichlet is expected to be more conservative in estimating the posterior $P_G(\lambda \mid \overrightarrow{s})$ over λ than the N-POPP model.

9.3 Evaluation on Synthetic Data

The previous chapter has shown that the N-POPP model can outperform both the POPP model and the POPP-Beta model in estimating the λ parameter of a Poisson process. As the

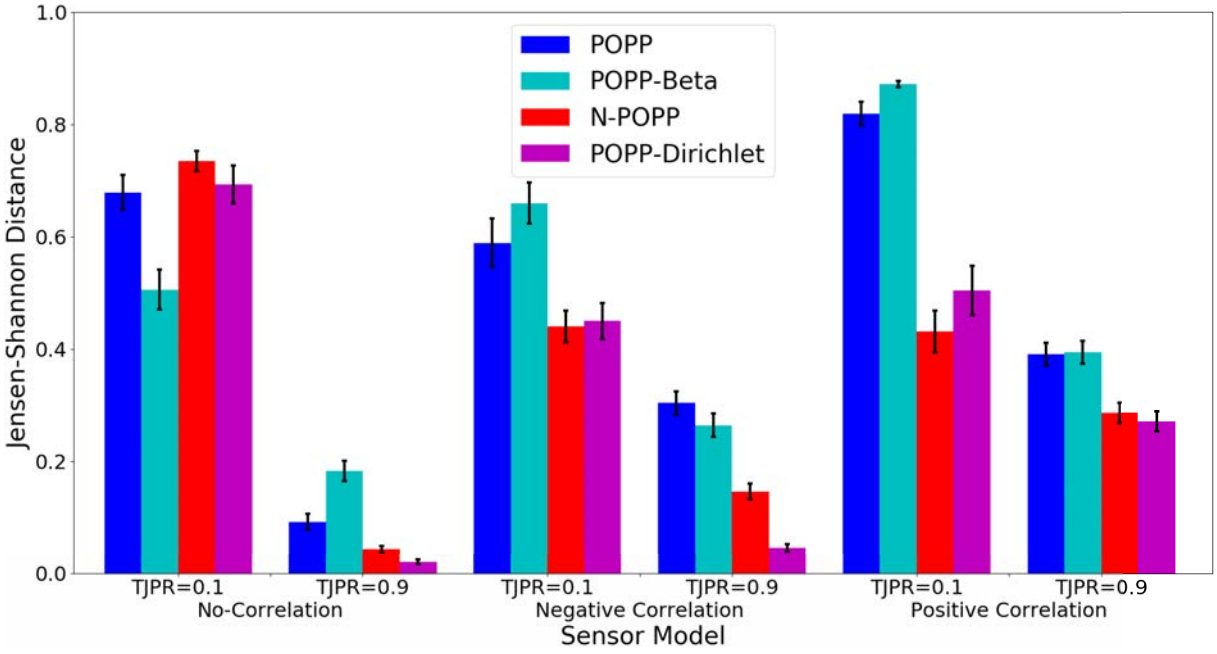

Figure 9.3: The Jensen-Shannon distance of posterior estimates of λ for the POPP-Dirichlet and other POPP models with 120 sample data used to build the (joint) sensor model with variation on $\mathscr{E}^+$. Each trial consisted of a stream of $\overrightarrow{s_1} \ldots \overrightarrow{s_{144}}$ samples to update $P_G(\lambda \mid \overrightarrow{s_i})$. Each data point is an average of 30 trials. Standard errors are shown.

POPP-Dirichlet is a direct extension of the N-POPP model, the simulated dataset used to evaluate the N-POPP model was borrowed. A comparison of the POPP-Dirichlet, the POPP model and its extensions was conducted. For this evaluation, the switching filter technique was chosen as the filter for all POPP models and its extensions, including the POPP-Dirichlet model. The uncorrected estimate, according to the FOPP model, was not included in this evaluation since its performance against N-POPP was shown in the previous chapter.

The joint sensor model was first built based on the sampled counts $x_1, \ldots, x_n$ from a Poisson process $P(x \mid \lambda' = 3)$ together with their corresponding sensor readings. Another set of counts $x_1, \ldots, x_n$ were sampled from the same process. These counts were then fed to a simulated sensor that counted unreliably, producing sensor readings $\overrightarrow{s_1} \ldots \overrightarrow{s_{144}}$. A recursive update, then, takes place on $P_G(\lambda \mid \overrightarrow{s_i})$ using the switching filter.

Two different sample sizes used to build the joint sensor model were chosen: a small number of samples, and a large number of samples. A small number of samples build an erroneous joint sensor model. In the POPP-Dirichlet model, a small number of samples

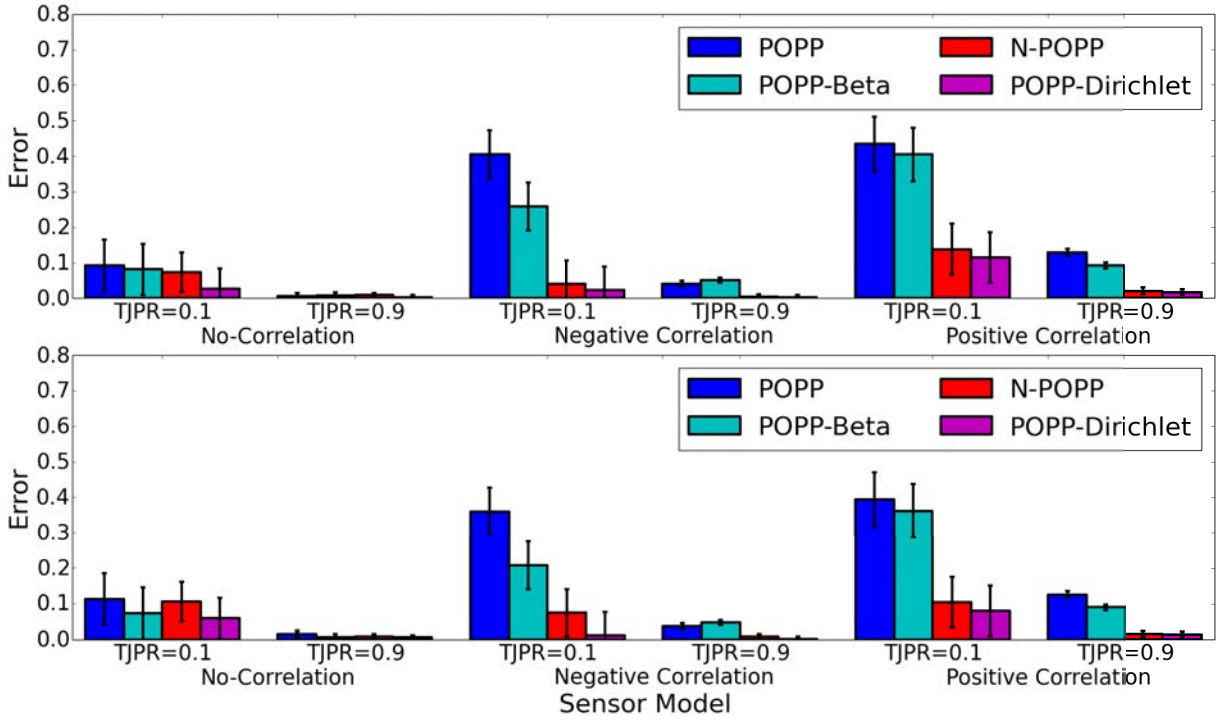

Figure 9.4: The RMSE of posterior estimates of λ for the POPP-Dirichlet and other POPP models with 2880 sample data used to build the (joint) sensor model with variation on $\mathscr{E}^+$. Each trial consisted of a stream of $\overrightarrow{s_1} \ldots \overrightarrow{s_{144}}$ samples to update $P_G(\lambda \mid \overrightarrow{s_i})$. Accuracies of MAP estimates are in the top panel, accuracies of expectation of the posterior in the bottom panel. Each data point is an average of 30 trials. Standard errors are shown.

creates a loose Dirichlet density. Conversely, a large number of samples creates a low variance, and thus reliable, joint sensor model by tightening the Dirichlet prior of the POPP-Dirichlet model. 120 samples were set for the small sample size, and 2880 samples were set for the large sample size.

Different correlations between two sensors were tested: positive correlation, negative correlation, and no correlation. For each correlation type, a further variation to different levels of sensor unreliability was considered. First, two variations were made to the true joint positive rate $\mathscr{E}^+$ (TJPR), while fixing the true joint negative rate $\mathscr{E}^-$ (TJNR) on each type of correlation. This includes:

- $P_{jnt}(d_{1k} = 1, d_{2k} = 1 \mid e_k = 1) = 0.1, P_{jnt}(d_{1k} = 0, d_{2k} = 0 \mid e_k = 1) = 0.9$ ($\mathscr{E}^+$ with low positive correlation);

- $P_{jnt}(d_{1k} = 1, d_{2k} = 1 \mid e_k = 1) = 0.9, P_{jnt}(d_{1k} = 0, d_{2k} = 0 \mid e_k = 1) = 0.1$ ($\mathscr{E}^+$ with high

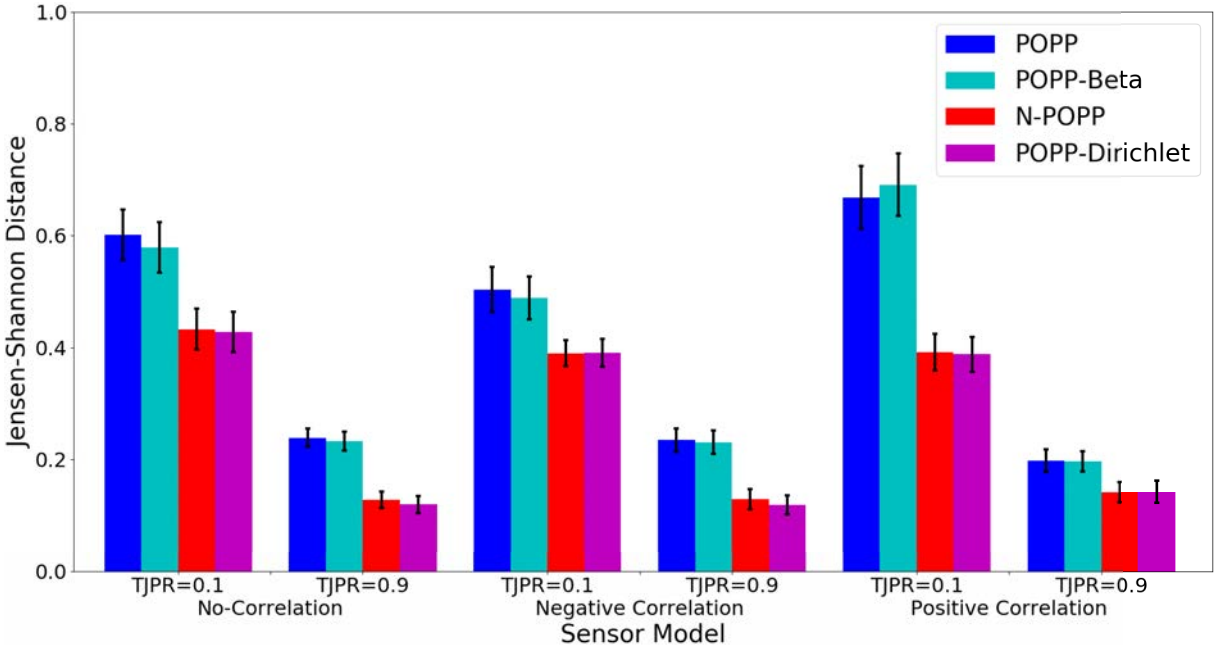

Figure 9.5: The Jensen-Shannon distance of posterior estimates of λ for the POPP-Dirichlet and other POPP models with 2880 sample data used to build the (joint) sensor model with variation on $\mathscr{E}^+$. Each trial consisted of a stream of $\overrightarrow{s_1} \ldots \overrightarrow{s_{144}}$ samples to update $P_G(\lambda \mid \overrightarrow{s_i})$. Each data point is an average of 30 trials. Standard errors are shown.

positive correlation);

- $P_{jnt}(d_{1k} = 1, d_{2k} = 0 \mid e_k = 1) = 0.05, P_{jnt}(d_{1k} = 0, d_{2k} = 1 \mid e_k = 1) = 0.05, P_{jnt}(d_{1k} = 0, d_{2k} = 0 \mid e_k = 1) = 0.9$ ($\mathscr{E}^+$ with low negative correlation);

- $P_{jnt}(d_{1k} = 1, d_{2k} = 0 \mid e_k = 1) = 0.45, P_{jnt}(d_{1k} = 0, d_{2k} = 1 \mid e_k = 1) = 0.45, P_{jnt}(d_{1k} = 0, d_{2k} = 0 \mid e_k = 1) = 0.1$ ($\mathscr{E}^+$ with high negative correlation);

- $P_{jnt}(d_{1k} = 1, d_{2k} = 0 \mid e_k = 1) = 0.033, P_{jnt}(d_{1k} = 0, d_{2k} = 1 \mid e_k = 1) = 0.033, P_{jnt}(d_{1k} = 1, d_{2k} = 1 \mid e_k = 1) = 0.033, P_{jnt}(d_{1k} = 0, d_{2k} = 0 \mid e_k = 1) = 0.901$ ($\mathscr{E}^+$ with no correlation – Similar to a sensor model with low TPR);

- $P_{jnt}(d_{1k} = 1, d_{2k} = 0 \mid e_k = 1) = 0.3, P_{jnt}(d_{1k} = 0, d_{2k} = 1 \mid e_k = 1) = 0.3, P_{jnt}(d_{1k} = 1, d_{2k} = 1 \mid e_k = 1) = 0.3, P_{jnt}(d_{1k} = 0, d_{2k} = 0 \mid e_k = 1) = 0.1$ ($\mathscr{E}^+$ with no correlation – Similar to a sensor model with moderate TPR).

Second, two variations to the true joint negative rate $\mathscr{E}^-$ (TJNR), while fixing the true joint positive rate $\mathscr{E}^+$ (TJPR) on each type of correlation. This includes:

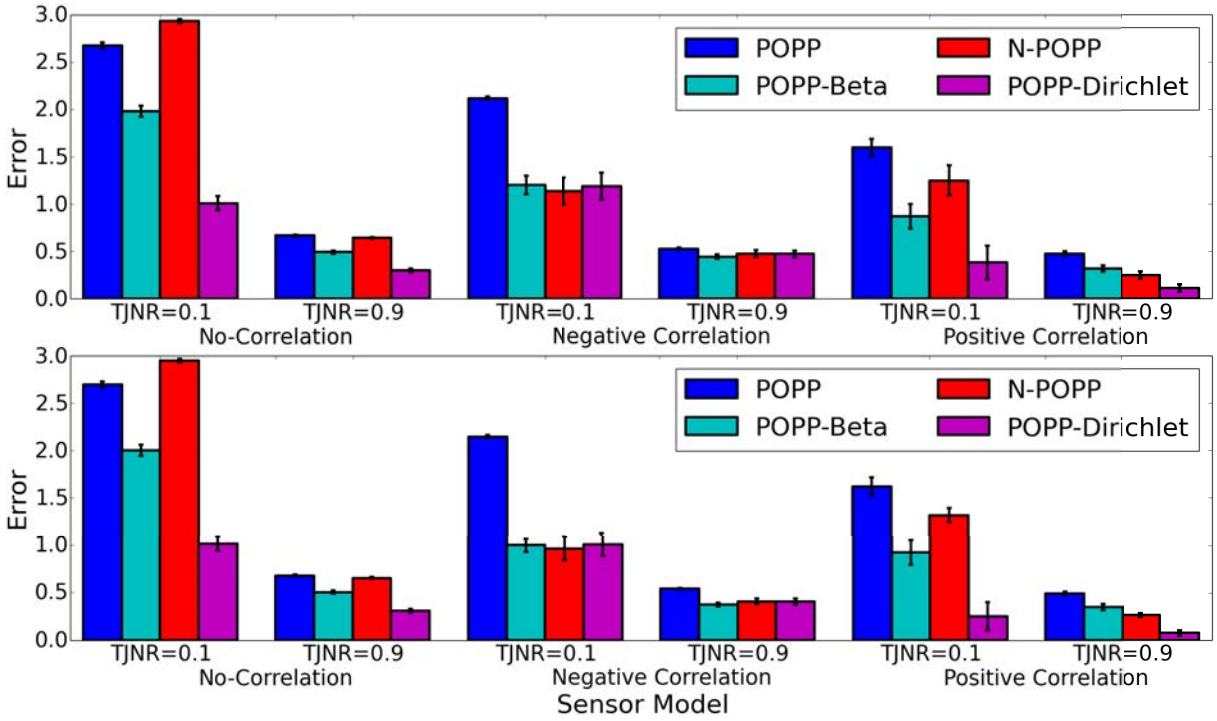

Figure 9.6: The RMSE of posterior estimates of λ for the POPP-Dirichlet and other POPP models with 120 sample data used to build the (joint) sensor model with variation in $\mathscr{E}^-$. Each trial consisted of a stream of $\overrightarrow{s_1} \ldots \overrightarrow{s_{144}}$ samples to update $P_G(\lambda \mid \overrightarrow{s_i})$. Accuracies of MAP estimates are in the top panel, accuracies of the expectation of the posterior in the bottom panel. Each data point is an average of 30 trials. Standard errors are shown.

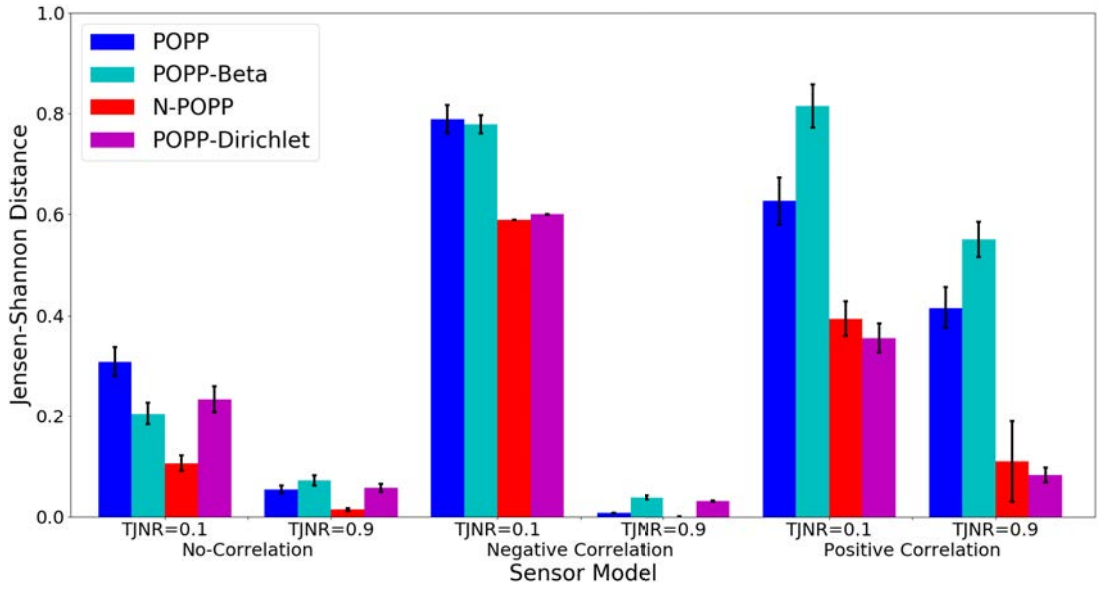

Figure 9.7: The Jensen-Shannon distance of posterior estimates of λ for the POPP-Dirichlet and other POPP models with 120 sample data used to build the (joint) sensor model with variation on $\mathscr{E}^-$. Each trial consisted of a stream of $\overrightarrow{s_1} \ldots \overrightarrow{s_{144}}$ samples to update $P_G(\lambda \mid \overrightarrow{s_i})$. Each data point is an average of 30 trials. Standard errors are shown.

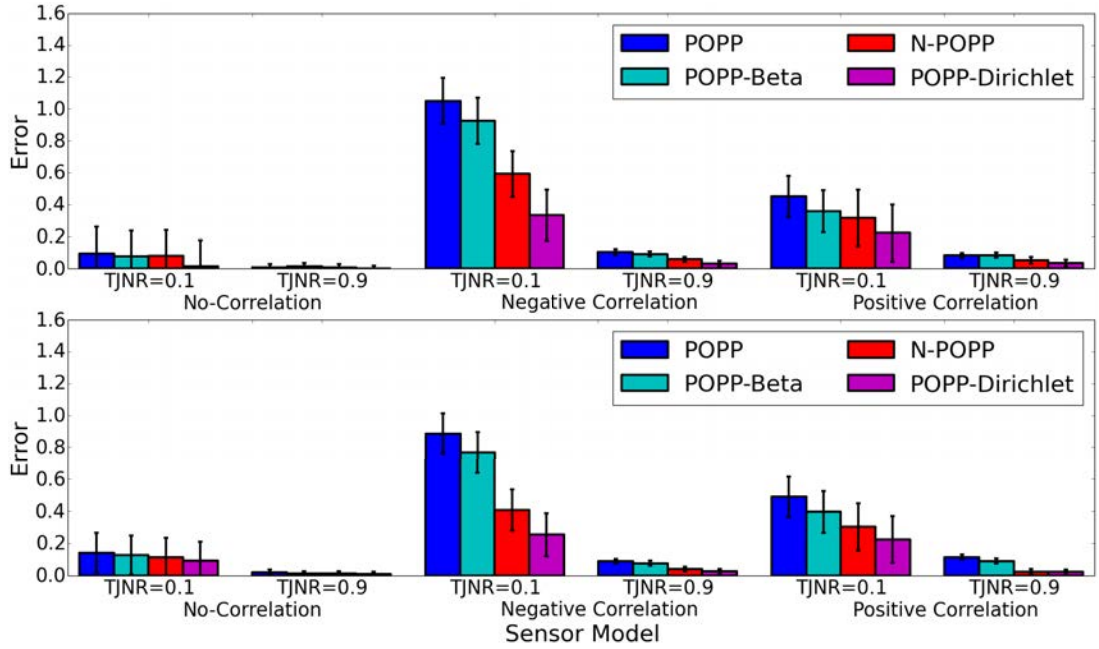

Figure 9.8: The RMSE of posterior estimates of λ for the POPP-Dirichlet and other POPP models with 2880 sample data used to build the (joint) sensor model with variation in $\mathscr{E}^-$. Each trial consisted of a stream of $\overrightarrow{s_1} \ldots \overrightarrow{s_{144}}$ samples to update $P_G(\lambda \mid \overrightarrow{s_i})$. Accuracies of MAP estimates are in the top panel, accuracies of the expectation of the posterior in the bottom panel. Each data point is an average of 30 trials. Standard errors are shown.

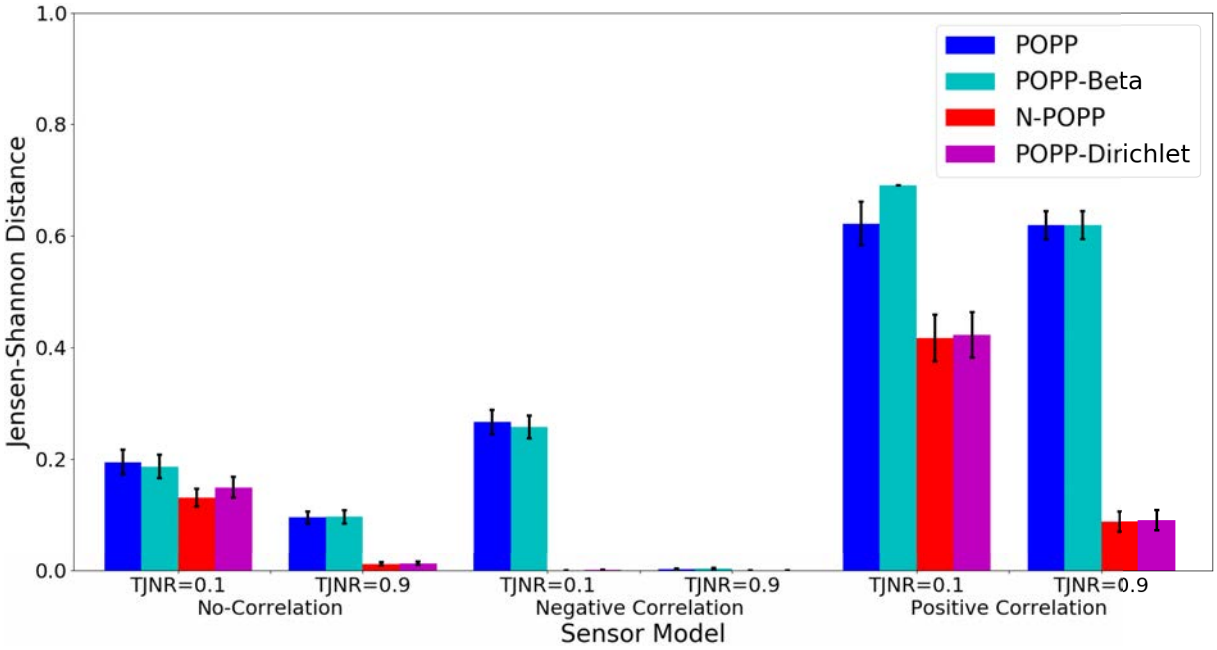

Figure 9.9: The Jensen-Shannon distance of posterior estimates of λ for the POPP-Dirichlet and other POPP models with 2880 sample data used to build the (joint) sensor model with variation on $\mathscr{E}^-$. Each trial consisted of a stream of $\overrightarrow{s_1} \ldots \overrightarrow{s_{144}}$ samples to update $P_G(\lambda \mid \overrightarrow{s_i})$. Each data point is an average of 30 trials. Standard errors are shown.

- $P_{jnt}(d_{1k} = 1, d_{2k} = 1 \mid e_k = 0) = 0.1, P_{jnt}(d_{1k} = 0, d_{2k} = 0 \mid e_k = 0) = 0.9$ ($\mathscr{E}^-$ with low positive correlation);

- $P_{jnt}(d_{1k} = 1, d_{2k} = 1 \mid e_k = 0) = 0.9, P_{jnt}(d_{1k} = 0, d_{2k} = 0 \mid e_k = 0) = 0.1$ ($\mathscr{E}^-$ with high positive correlation);

- $P_{jnt}(d_{1k} = 1, d_{2k} = 0 \mid e_k = 0) = 0.05, P_{jnt}(d_{1k} = 0, d_{2k} = 1 \mid e_k = 0) = 0.05, P_{jnt}(d_{1k} = 0, d_{2k} = 0 \mid e_k = 0) = 0.9$ ($\mathscr{E}^-$ with low negative correlation);

- $P_{jnt}(d_{1k} = 1, d_{2k} = 0 \mid e_k = 0) = 0.45, P_{jnt}(d_{1k} = 0, d_{2k} = 1 \mid e_k = 0) = 0.45, P_{jnt}(d_{1k} = 0, d_{2k} = 0 \mid e_k = 0) = 0.1$ ($\mathscr{E}^-$ with high negative correlation);

- $P_{jnt}(d_{1k} = 1, d_{2k} = 0 \mid e_k = 0) = 0.033, P_{jnt}(d_{1k} = 0, d_{2k} = 1 \mid e_k = 0) = 0.033, P_{jnt}(d_{1k} = 1, d_{2k} = 1 \mid e_k = 0) = 0.033, P_{jnt}(d_{1k} = 0, d_{2k} = 0 \mid e_k = 1) = 0.901$ ($\mathscr{E}^-$ with no correlation – Similar to a sensor model with low TNR);

- $P_{jnt}(d_{1k} = 1, d_{2k} = 0 \mid e_k = 0) = 0.3, P_{jnt}(d_{1k} = 0, d_{2k} = 1 \mid e_k = 0) = 0.3, P_{jnt}(d_{1k} = 1, d_{2k} = 1 \mid e_k = 0) = 0.3, P_{jnt}(d_{1k} = 0, d_{2k} = 0 \mid e_k = 1) = 0.1$ ($\mathscr{E}^-$ with no correlation – Similar to a sensor model with moderate TNR).

The performance of the POPP-Dirichlet model was assessed in a similar manner to previous chapters by comparing the RMSE of the two estimators and the Jensen-Shannon distance of each posterior over λ to the true λ'. The result of the assessment when TJPR $\mathscr{E}^+$ was varied is shown by Figure 9.2 and 9.3 for a low number of samples and Figures 9.4 and 9.5 for a large number of samples. The result of the assessment when TJNR $\mathscr{E}^-$ was varied is shown by Figure 9.6 for a low number of samples and Figure 9.8 and 9.9 for a large number of samples. In the RMSE assessment, all the figures clearly show that the POPP-Dirichlet model outperformed any of the POPP model variations from previous chapters. This becomes more apparent when there is only a small number of samples available. Figures 9.2 and 9.6 clearly show that, with a low number of samples, any POPP model which models the unreliability of a sensor model, such as the POPP-Beta and the POPP-Dirichlet models, provides a better estimate than their counterpart which does not model the unreliability of the sensor model.

In the Jensen-Shannon distance assessment, Figures 9.5 and 9.9 show that the POPP-Dirichlet model has comparable performance to the N-POPP model especially when there is a large number of samples to train the joint sensor models. However, the same statement cannot be made when there is only a small number of samples available to train the joint sensor models of the POPP-Dirichlet and the N-POPP models, although in general, the performance of those models are still better than the POPP and the POPP-Beta models.

9.4 Evaluation on A Real World Dataset

Once again, in a similar fashion to the last three chapters, the POPP-Dirichlet filter is compared with the POPP filter and two of its extensions (the POPP-Beta filter and the N-POPP filter) in estimating the λ parameter of both a homogeneous Poisson process and a periodic Poisson process on a real world dataset. The POPP-Dirichlet filter is also compared with the FOPP filter. Similar to the evaluation on the simulated data set, the switching filter is used for all the models except the standard FOPP model.

The same dataset used in the last three chapters is used. Available perception algorithms attached to a mobile robot are leg detector (LD), upper body detector (UBD), and change detector (CD). Each minute, each detector returns a binary value whether there was a person or not. All binary detections coming from the same minute are grouped together, and these groups are further clustered into a 10-minute cluster to form a joint sensed count of the number of people all sensors detect in each 10 minute interval during the day. The detections are still organised according to time/date and the spatial region where each detection was made.

Both the joint sensor model and the standard sensor model must be trained from sensor counts and true counts. The sensor models were built from the data collected from 10am-8pm each for 48 days. A four fold cross-validation on the 48-day dataset with the unit being a whole day was carried out, where 12 days of data were used to train the sensor model. This process is done for all areas within the patrol space. As a reminder, an example of a joint sensor model

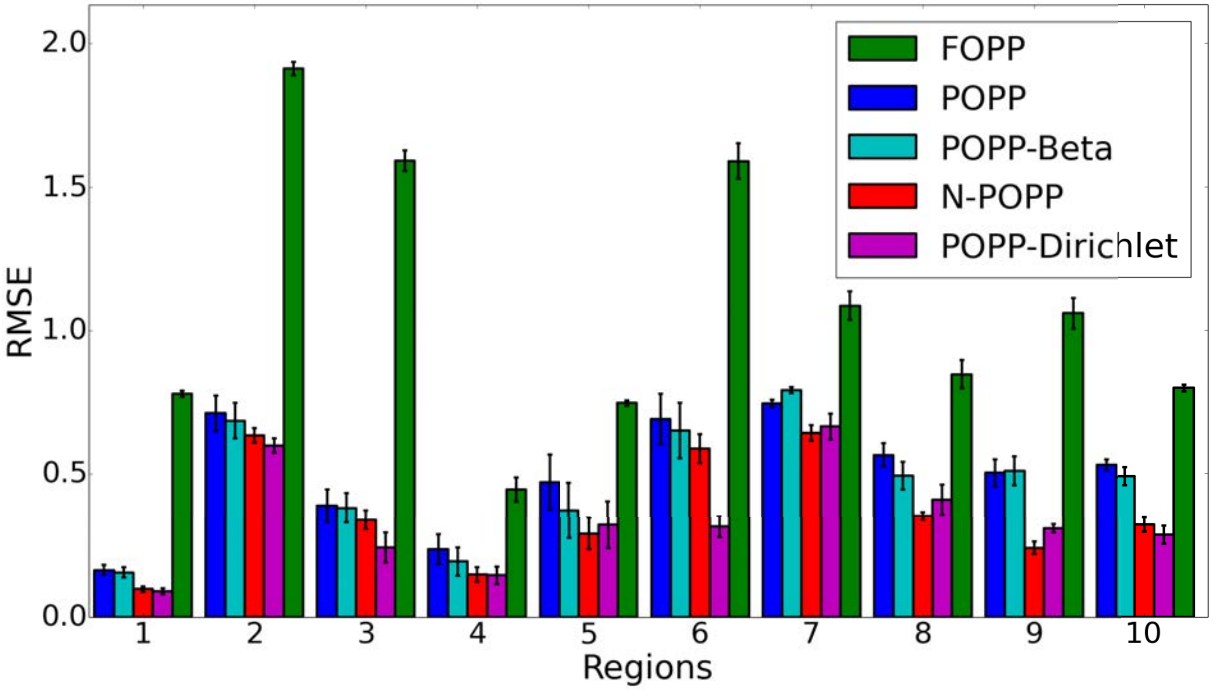

Figure 9.10: The RMSE of the FOPP, POPP, POPP-Beta, and POPP-Dirichlet estimators of λ as it varies across areas (regions) of the environment. Standard error is shown.

for the N-POPP filter trained from 48 days of data can be seen in Table 8.1 (in the previous chapter).

The true λ for the Poisson process and $\lambda(t_i, t_j)$ for the periodic Poisson process on each region was estimated by running a FOPP filter on the true counts. The uncorrected estimate λ according to the FOPP model was estimated only from the change detector count data since the change detector is the most reliable detector among three detectors available. Using RMSE and the Jensen-Shannon distance, the estimated true λ and its distribution are then used as a comparison to the resulting estimated λ (and $\lambda(t_i, t_j)$ respectively) produced by each filter.

9.4.1 The POPP-Dirichlet models on Homogeneous Poisson Processes

This section starts with a case on homogeneous Poisson processes with an interval of 10 minute. The results are shown in Figure 9.10 for the RMSE and 9.11 for the Jensen-Shannon distance. These figures are an extension of Figure 8.9 and 8.10 by adding an evaluation of

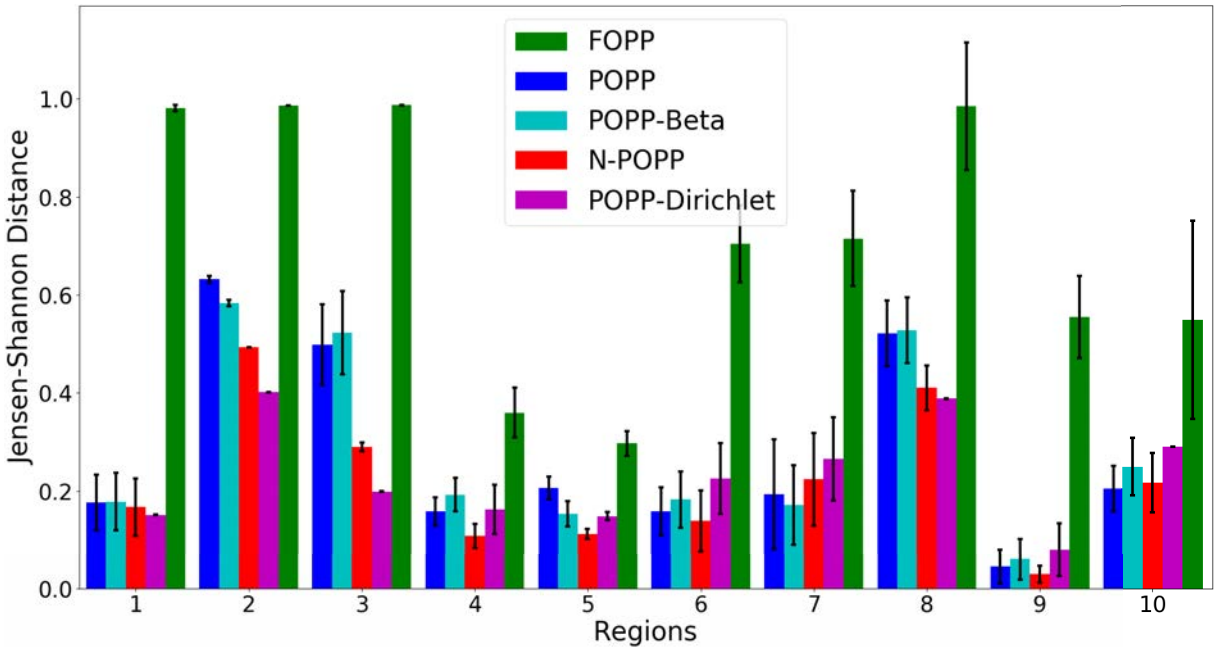

Figure 9.11: The Jensen-Shannon distance of the FOPP, the POPP, the POPP-Beta, the N-POPP, and the POPP-Dirichlet model distributions of λ as it varies across areas (regions) of the environment. Standard error is shown.

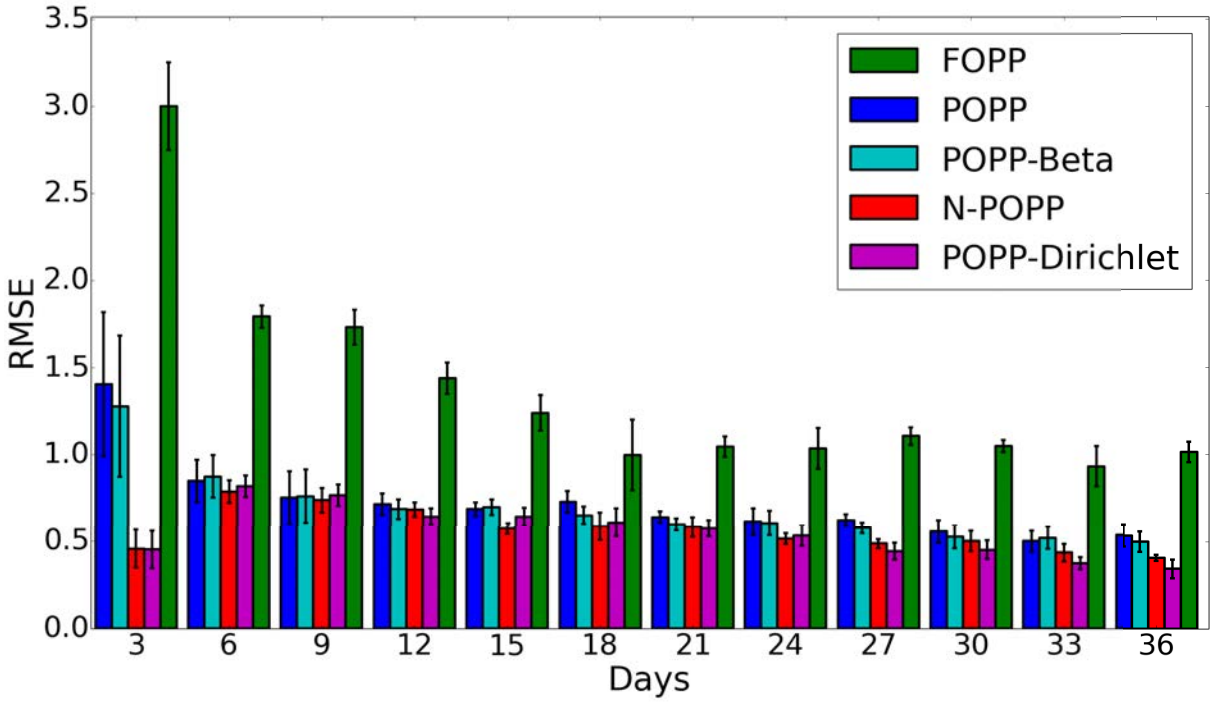

Figure 9.12: The RMSE evolution from day 3 to day 36 with 3 day interval, averaged across all regions. Standard error is shown.

the POPP-Dirichlet filter. In terms of RMSE, the λ estimate produced by the POPP-Dirichlet filter is more accurate than the ones produced by the standard POPP filter and the POPP-Beta

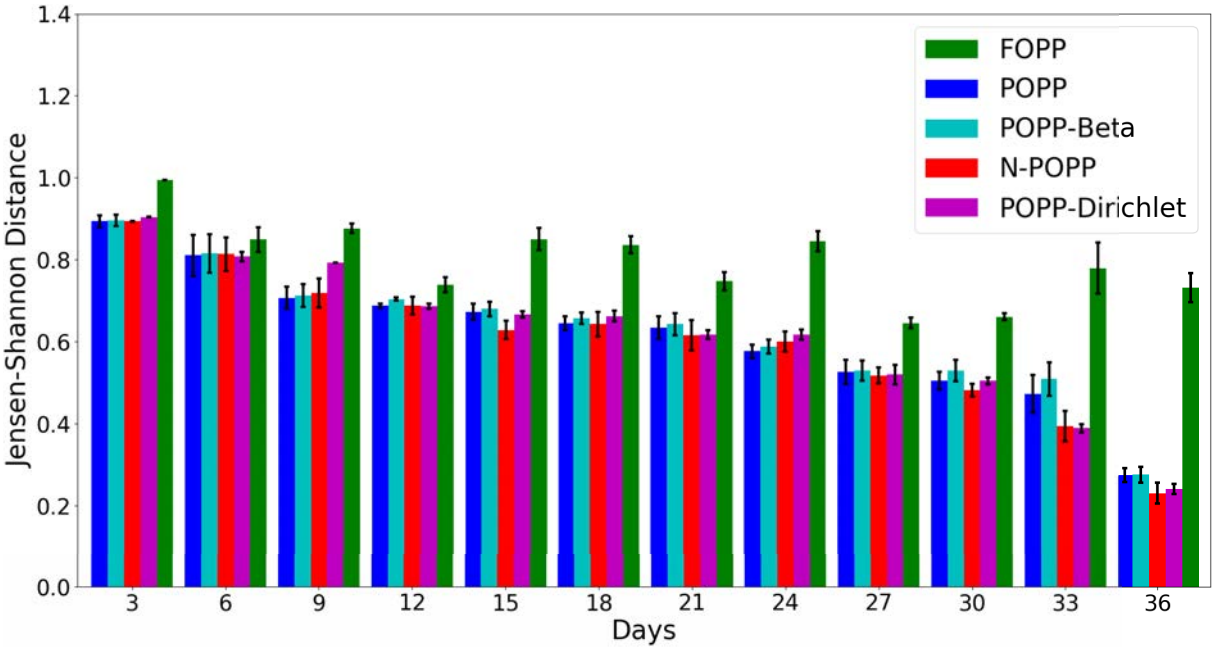

Figure 9.13: The Jensen-Shannon distance evolution of the FOPP, the POPP, the POPP-Beta, the N-POPP, and the POPP-Dirichlet model distributions of λ from day 3 to day 36 with 3 day interval, averaged across all regions. Standard error is shown.

filter. However, the estimate is not always more accurate compared to the one produced by the N-POPP filter. As the POPP-Dirichlet is more conservative in estimating the parameter λ, the estimate moves rather slowly towards the true λ and the distribution produced by the POPP-Dirichlet tends to be wider than the N-POPP model. This argument is showcased in Figures 9.11 where the Jensen-Shannon distance shows that the N-POPP model produced more similar distributions to the true λ than the POPP-Dirichlet.

Similar to the last three chapters, an evaluation on how the POPP-Dirichlet evolved with time was conducted in terms of both RMSE and Jensen-Shannon distance, since it is expected to gradually get closer to the true λ. Figures 9.12 and 9.13 show that as time passes the performance of the POPP-Dirichlet estimator becomes better and slowly outperforms other estimators including the N-POPP estimator.

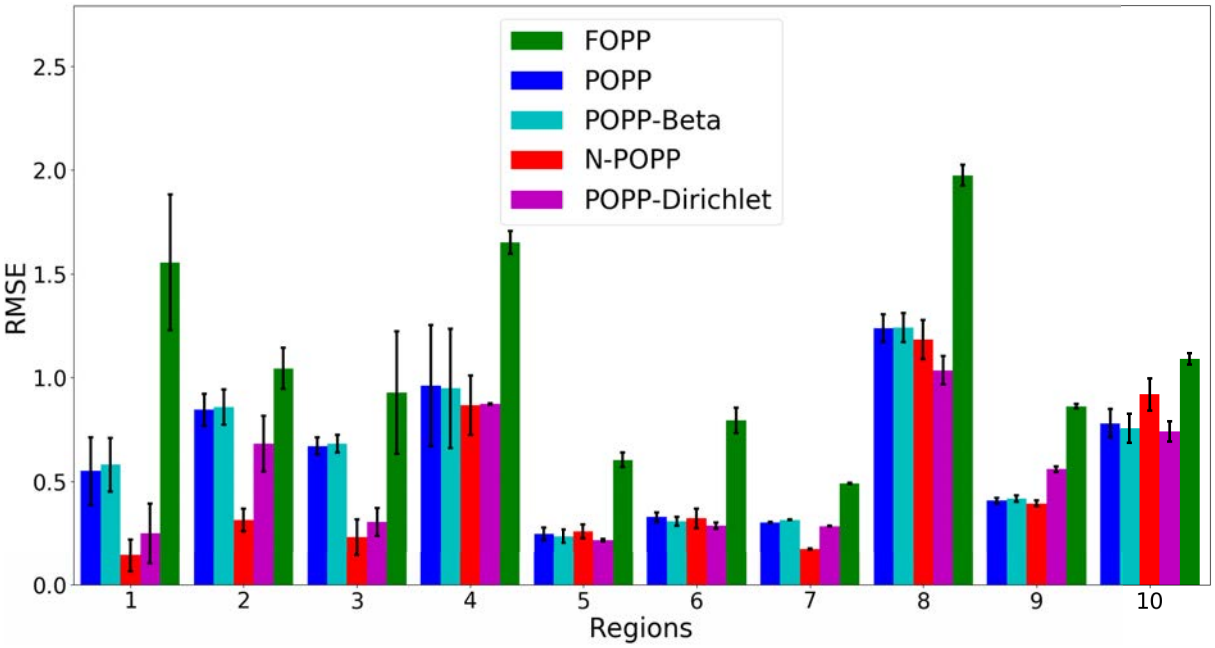

Figure 9.14: The RMSE of the FOPP, POPP, POPP-Beta, N-POPP, and POPP-Dirichlet estimators of $\lambda(t_i, t_j)$ as it varies across areas (regions) of the environment. Standard error is shown.

9.4.2 The POPP-Dirichlet models on Periodic Poisson Processes

This evaluation, a periodic Poisson process in which the imposed single periodicity is a one-day cycle, i.e. $\lambda(t_i, t_j) = \lambda(t_{i+\Delta}, t_{j+\Delta})$ with $\Delta = 24 * 60$(minutes), is assumed to be the underlying process. The results are shown in Figure 9.14 for the RMSE and 9.15 for the Jensen-Shannon distance.

Similar to the N-POPP filter, the POPP-Dirichlet filter is able to cope and overcome the problems with limited sample data both for building the joint sensor model and estimating the $\lambda(t_i, t_j)$. In many regions, the POPP-Dirichlet managed to show better estimates as well as more similar distributions than the POPP, the POPP-Beta, and the FOPP filters. However, the POPP-Dirichlet filter falls behind both in accuracy (RMSE) and distribution similarity compared to the N-POPP filter. This is attributed to the POPP-Dirichlet conservative way in estimating the parameter $\lambda(t_i, t_j)$ compared to the N-POPP filter.

Once again, a side evaluation on how the POPP-Dirichlet evolved with time was evaluated. Figures 9.16 and 8.16 show the evolution of each filter overtime in terms of RMSE and Jensen-Shannon distance across regions. The POPP-Dirichlet filter gradually increased its ac-

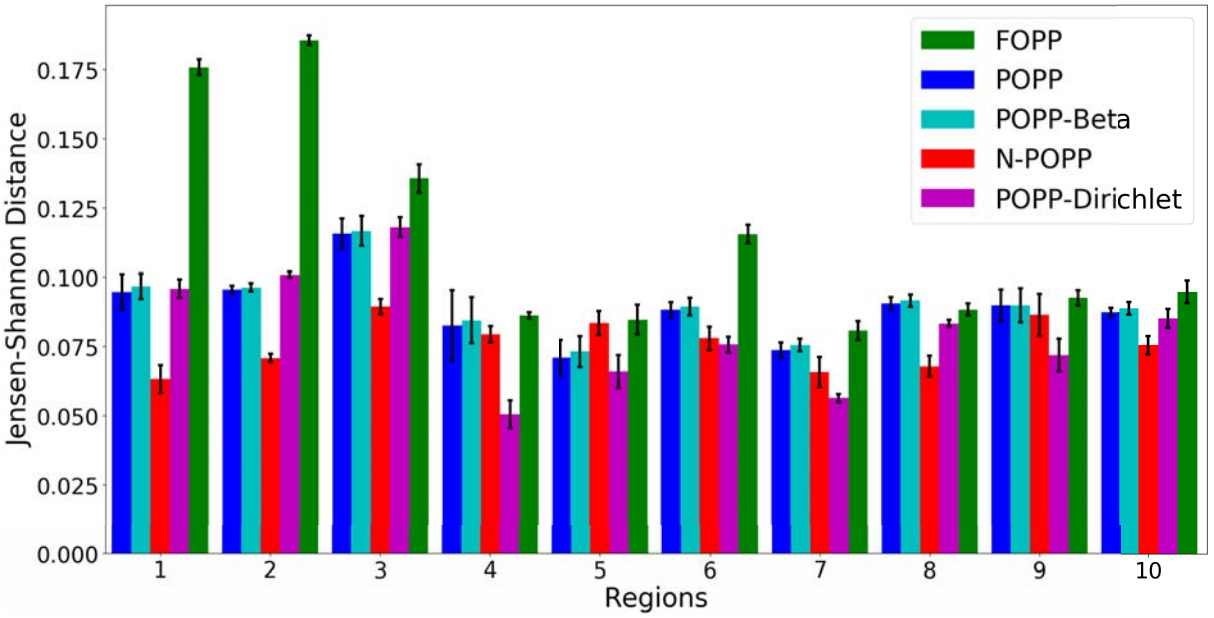

Figure 9.15: The Jensen-Shannon distance of the FOPP, POPP, POPP-Beta, N-POPP, and POPP-Dirichlet model distributions of $\lambda(t_i, t_j)$ as it varies across areas (regions) of the environment. Standard error is shown.

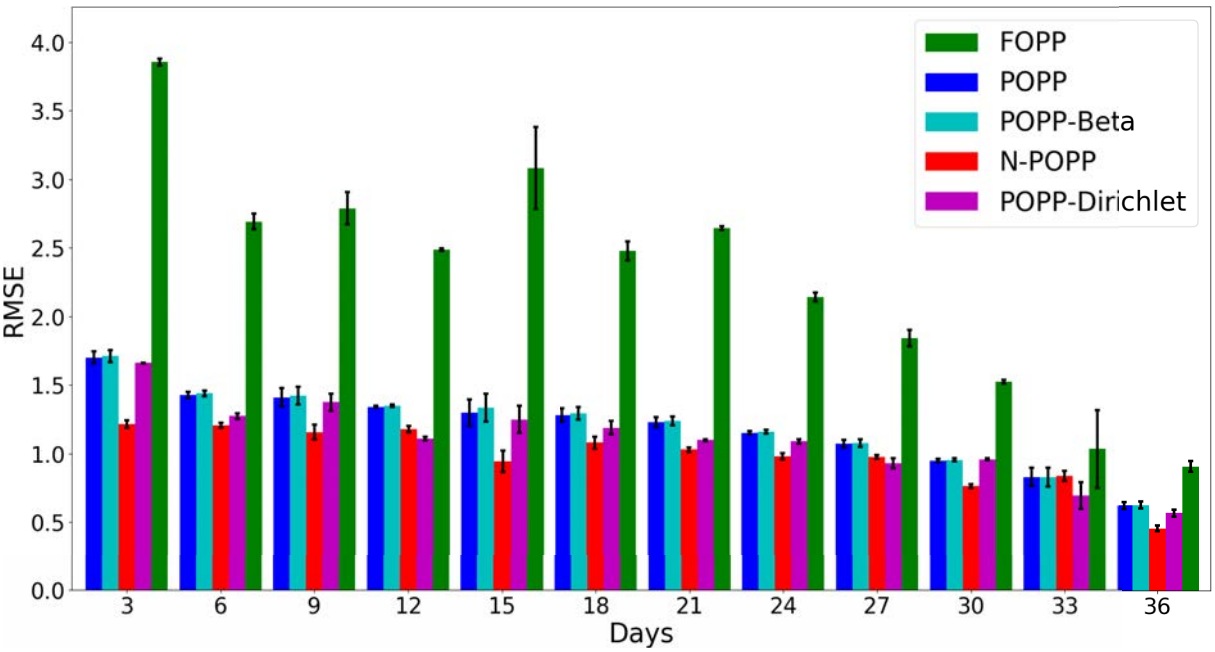

Figure 9.16: The RMSE evolution of periodic Poisson processes with POPP, POPP-Beta, N-POPP, POPP-Dirichlet and FOPP filters from day 3 to day 36, averaged across all regions. Standard error is shown.

curacy in estimating $\lambda(t_i, t_j)$ overtime. The filter outperformed the POPP and the POPP-Beta filters in terms of accuracy. Unlike what is shown in Figure 9.12, the POPP-Dirichlet filter here was a bit outperformed by the N-POPP filter in terms of RMSE accuracy.

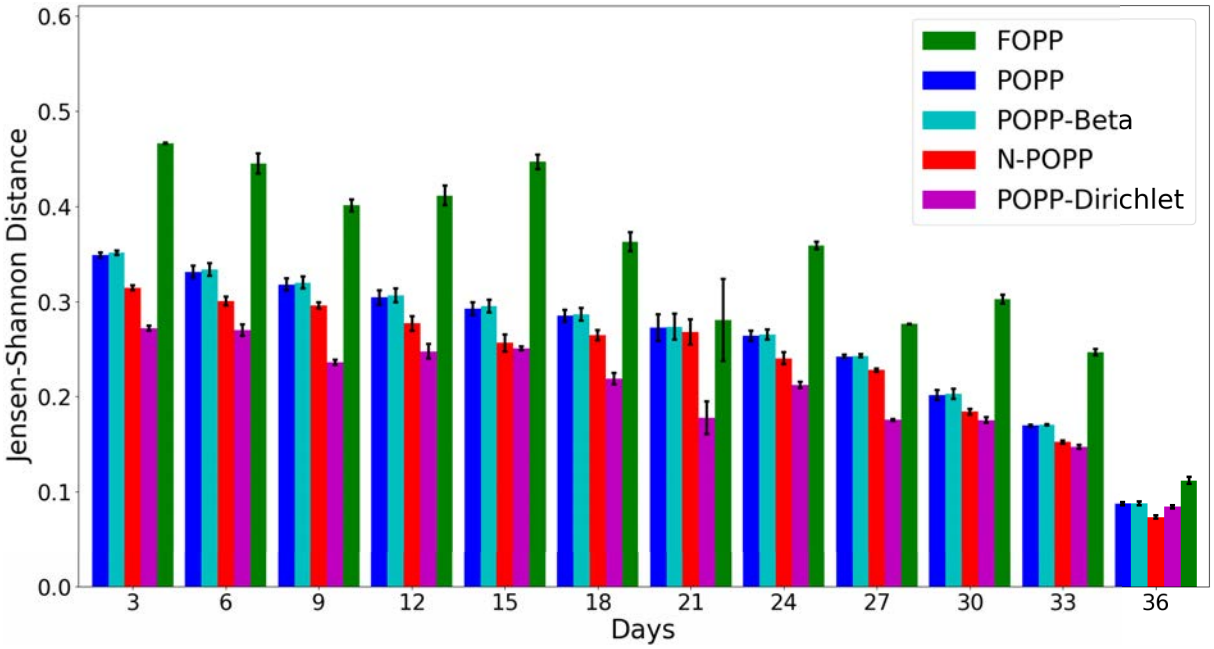

Figure 9.17: The Jensen-Shannon distance evolution of the FOPP, the POPP, the POPP-Beta, the N-POPP, and the POPP-Dirichlet filters in periodic Poisson processes from day 3 to day 36 in a 3-day interval, averaged across all regions. Standard error is shown.

9.5 Discussion

In this chapter, another extension–POPP-Dirichlet–to the POPP model was explained and tested. POPP-Dirichlet is, loosely, a combination of two previous extensions of the POPP model: the POPP-Beta and the N-POPP. By merging the idea of two previous extensions, the POPP-Dirichlet does not have either limitation that the POPP model has. The POPP-Dirichlet is based on the N-POPP model where the joint sensor model is modeled as latent variables which follow Dirichlet densities. An analytic closed form Dirichlet-multinomial distribution provides a replacement of the multinomial distribution in Equation 9.2.

The experiments on simulated data showed that the POPP-Dirichlet estimator outperformed any other estimator in estimating the parameter λ of a Poisson process. Because the POPP-Dirichlet is more conservative than its counterpart, the N-POPP model, its estimator moves rather slowly in estimating the parameter λ. This is clearly shown by the evolution of the accuracy of the estimator on the real world data. This becomes more apparent whenever only limited sample data are available.

CHAPTER 10

EXPLORE-EXPLOIT WHEN OBSERVING HUMAN ACTIVITIES

In the first part of this thesis, the focus was about a robot that learns the level of aggregate human activities over an extended period. It was argued that in order to achieve this, regardless of specific learning models, the robot must go to places where humans are bound to be found. Up to this point, the thesis has been concerned with Bayesian methods for correctly estimating how many humans might be encountered in a certain place at a particular time given that available sensors in the robot are unreliable. In this chapter, the work from the first part of this thesis (spectral-Poisson models for extracting periodic structures on temporally sparse count data) is brought together with the second part of this thesis (Bayesian inference for partially observable Poisson processes). This chapter shows how to use the resulting predictions to drive exploration for human activities. In particular, the aim is to drive the robot to go where there are the most people performing activities, i.e. where the aggregate level of human activity is highest. This leads to solving an exploration-exploitation trade-off: the robot must explore to find out where humans congregate, so as to then exploit that by observing as many human activities as possible in a limited time.

This chapter starts with an introduction to why exploring to maximise observations of human activities can be thought of as an exploration-exploitation problem which is well known in the reinforcement learning community. As a proof of concept, an evaluation of the spectral-Poisson model, introduced in Chapter 4, in driving a mobile robot to explore and observe the Aruba dataset is introduced. This dataset provides ground truth data for measuring the performance of the spectral-Poisson model in simulating a robot exploration. As a comparison to the spectral-Poisson, the periodic Poisson model, described in 4.1, and a random exploration model are included. At the final section, the spectral-POPP models and the POPP filters, which have been the focus of Chapter 6 to 9, are brought together to drive exploration by a mobile robot for a series of nine-week deployments. Altogether, these two different statistical methods give the ability to leverage the periodic structure of human behaviour to perform better exploration.

10.1 Formulating Exploration-Exploitation Problems

In Chapter 4, the spectral-Poisson model has shown its ability to retrieve long-term, re-occurring patterns from a time series of count data. Given that the underlying process of interest has periodic structures, the spectral-Poisson predicts the fluctuation in the count data at particular times and places. This approach was integrated as part of an overall mobile robot system. This robot was employed in a "security scenario" in several office buildings. In this scenario, the robot actively planned to observe places with a predicted high-level of aggregate activity, so as to gather as much data on human activities as possible.

The problem to plan places to observe which will give the highest number of human activities can be seen as an exploration-exploitation problem. Exploration-exploitation problems arise when an agent, in this scenario the mobile robot acts as the agent, does not fully understand the process it is trying to control. In any time, the robot is given two options. It can either spend its time and resources to better understand the process (explore) for a better action later on, but sacrificing its short-term reward, or spend its time and resources to

exploit what the robot already understands to gain reward and risk permanently following its policy that might be suboptimal. In each time the robot has a choice between many actions, each of which both explores and exploits a certain place, but to varying degrees. As its goal is to maximise the reward gathered–as in getting as much data on human activities as possible–given its limited operational life, it is preferable to have a policy that is as near optimal as possible.

For example, let's assume that this is 12 o'clock on a workday in an office with a lunch room. The robot can choose between the corridor, where it has seen an average of three people at this time , and the lunch room, which it has never visited. Given the prior over the average of human activities in both places, one option to choose is to visit the lunch room which might turn up having a lot of people at that time.

This problem is well known in the reinforcement learning community, and it is known as the problem of optimising the exploration-exploitation trade-off. By mapping the number of human activities the robot can find in a place to a certain reward, the robot exploration is indeed a reinforcement learning problem in which the robot has an exploration-exploitation problem.

While exploration-exploitation problems in reinforcement learning are typically intractable, there are well known approximate answers that are very quick to compute Wyatt (1998), Alba & Dorronsoro (2005), Audibert et al. (2009). One such answer is to the use the upper bound of the probability interval over the reward. Here that means the upper bound of the probability interval of the aggregate number of activities. The other option, which is easier to compute, is to use the upper bound of the probability interval of the arrival rate (λ) of a Poisson process.

In this robot exploration, the latter option to use the upper bound of the probability interval of λ is chosen because it tackles both aleatoric uncertainty and epistemic uncertainty. While there is (aleatoric) uncertainty in the aggregate number of activities the robot will see in an area in a particular period, there is part of the (epistemic) uncertainty which is due to robot's lack of exploration in observing the area at that time. This kind of uncertainty is

Algorithm 4 *Spectral Upper Bound*

Input: $(\alpha_1, \beta_1), \ldots, (\alpha_n, \beta_n)$: Poisson process
Output: $\lambda_1^{ub}, \ldots, \lambda_n^{ub}$: spectral upper bound
Procedure:

1. Init. k $\leftarrow$ 1, m $\leftarrow \eta$
2. Repeat until k > n
   - $k \leftarrow k+1$
   // Get the upper bound of the confidence interval
   - $\lambda_k \leftarrow CDF(0.95, \alpha_k, \beta_k)$

// Transform $\lambda_1, \ldots, \lambda_n$ to spectrums with l-AAM technique

3. $\mathscr{S} \leftarrow$ **Algorithm1**$(\lambda_1, \ldots, \lambda_n, \text{l})$
5. Init. k $\leftarrow$ 1, $\lambda_1^{ub}, \ldots, \lambda_n^{ub} \leftarrow (0, \ldots, 0)$
4. Repeat until k > m
   // Create a cosine signal from $\mathscr{S}[k]$
   $[\mid \omega_k \mid, arg(\omega_k), \omega_k] \leftarrow \mathscr{S}[\omega_k]$
   - $x_1, \ldots, x_n \leftarrow \mid \omega_k \mid * cos(2\pi * \omega_k + arg(\omega_k))$

   // Add current $\lambda_1^{ub}, \ldots, \lambda_n^{ub}$ with the cosine signal
   - $\lambda_1^{ub}, \ldots, \lambda_n^{ub} \leftarrow \lambda_1^{ub}, \ldots, \lambda_n^{ub} + x_1, \ldots, x_n$

not directly in the aggregate number of activities, but it is in λ, the parameter of the Poisson process.

This means that the Poisson and spectral-Poisson model, introduced in Chapter 4, use the upper bound of the confidence interval of the arrival rate (λ_{UB}) to provide places for the robot to visit. The upper bound of the confidence interval of the arrival rate of the non-homogeneous Poisson process is calculated as follows

$$\lambda_{UB}(t_i, t_j) = \int_{t_i}^{t_j} CDF^{-1}(\% = 0.95 \mid \alpha_t, \beta_t)\, dt \tag{10.1}$$

with $\lambda_{UB}(t_i, t_j)$ is the upper bound of λ within time t_i and t_j, $i, j \in \{1, \ldots, \Delta\}$, and CDF^{-1} is the inverse of the cummulative density function of a gamma distribution. In the case of a spectral-Poisson model, a series of upper bounds chosen for each $\lambda(t_i, t_j)$ were then encoded and extracted via spectral analysis with l-AAM technique to produce a smoother series of upper bound estimates $\lambda'_{UB}(t_i, t_j)$. Algorithm 4 depicts the process of choosing the upper bound of $\lambda(t_i, t_j)$ of a Poisson process and applying spectral analysis to it. Given a series of upper bounds $\lambda^r_{UB}(t_i, t_j)$ for each place, places which will be visited between time t_i and t_j is

chosen by

$$\arg\max_{r \in \mathscr{R}} \lambda^{r}_{UB}(t_i, t_j) \tag{10.2}$$

with $\mathscr{R}$ as a set of specified regions.

10.2 Evaluation on the Aruba Dataset

Before the robot was deployed as a security guard, an evaluation was conducted to assess how much improvement, concerning data gathering, the robot could obtain by replacing a random exploration policy with an upper bound exploration using the spectral-Poisson model. To test the (activity) exploration, the Aruba dataset was used. This dataset was first introduced by Cook in Cook (2010), and then modified by Coppola et al., in Coppola, Krajník, Duckett & Bellotto (2016). The 'Aruba' dataset contains ground-truth activities of a home-bound person in a small apartment for 16 weeks Cook (2010). These activities include: Bed to Toilet, Eating, Enter Home, Housekeeping, Leave Home, Meal Preparation, Relax, Resperate, Sleeping, Wash Dishes, Work. Coppola et al. modified the dataset by partitioning the small apartment into nine regions consisting second bathroom, junction, office, living room, master bedroom, master bathroom, corridor, second bedroom, kitchen Coppola, Krajník, Duckett & Bellotto (2016). Furthermore, it provides an estimated minute-by-minute person location based on the apartment's motion detectors. Thus, the 'Aruba' dataset contains a minute-by-minute timeline of 12 different activities performed at 9 different locations over the course of 16 weeks[1].

We simulated the exploration behaviour by creating an imaginary autonomous mobile robot equipped with an unreliable perception algorithm. The imaginary perception/sensor has a true positive rate of 0.8, and a true negative rate of 0.8. As a correction to the systematic error, such as applying the POPP filter or any of its extension, was not implemented in this simulation, it is not of interest to know whether there is an error (miscount) or not. Each day,

[1] https://github.com/gestom/fremen_activity_benchmark

Table 10.1: A Comparison of the average percentage of the total duration of positive observations over the total duration of activities taking place in that location. The exploration policies are: random exploration, the periodic Poisson process and the spectral-Poisson process.

Room	Random	periodic PP	spectral-Poisson
Second Bathroom	**11.66**%	9.09%	4.55%
Junction	**11.20**%	8.37%	4.35%
Office	**11.11**%	8.65%	4.86%
Living Room	10.66%	35.70%	**63.66**%
Master Bedroom	10.13%	56.04%	**79.21**%
Master Bathroom	**10.10**%	8.25%	3.55%
Corridor	**11.38**%	6.89%	2.01%
Second Bedroom	**10.21**%	7.32%	5.74%
Kitchen	**10.68**%	9.61%	4.01%

the imaginary robot planned a series of places to visit, which included the duration of the visit to each place, to look for possible human activities. The recommended places, and when the robot should visit those places, were given by an exploration model. With this plan, the robot then moved to the location, detected and recorded activities for the given time period. This process ran throughout the day and looped throughout the whole 16 weeks.

Three approaches were applied for the exploration model: the Poisson model, the spectral-Poisson model, and the Random exploration model. All these models were based on periodic Poisson processes with Δ set to a week period. The random model provided a random place for the robot to visit throughout the entire day. The Poisson and the spectral-Poisson models follow the exploration policy using the upper bound of the confidence interval of the arrival rate described in the previous section.

Recall that the focus of this evaluation on the simulated environment is to gather count data as much as possible while learning and predicting the regular patterns of the home-bound person. A comparison among different exploration strategies was made, based on how many observations the robot makes using a particular exploration model. One comparison is of the total duration of positive observations over the total duration of activities taking place in that location. This is shown by Table 10.1. The other comparison is of the total duration of positive observations over the total duration of the robots' visits. This is shown by Table 10.2.

Table 10.1 shows that for places with strong periodic patterns of activities such as living

Table 10.2: A Comparison of the average percentage of the total duration of positive observations over the total duration of the robots' visits to that location. The exploration policies are: random exploration, the periodic Poisson process and the spectral-Poisson process.

Room	Random	periodic PP	spectral-Poisson
Second Bathroom	14.94%	19.18%	**42.07**%
Junction	0.99%	1.25%	**1.40**%
Office	0.16%	0.16%	**0.19**%
Living Room	36.23%	54.35%	**57.86**%
Master Bedroom	33.45%	67.66%	**70.06**%
Master Bathroom	2.98%	4.37%	**4.62**%
Corridor	3.77%	**4.09**%	4.07%
Second Bedroom	2.11%	**2.68**%	2.29%
Kitchen	5.10%	**7.26**%	6.22%

room and master bedroom, the periodic Poisson model and the spectral-Poisson model were able to significantly improve the robot's presence during activity in the locations. However, both models performed worse than the random exploration if places do not have strong periodicity in the activities. In contrast to Table10.1, Table 10.2 shows that the spectral-Poisson is the most efficient model in finding activities. This is shown by its hit and miss ratio of observations. This supports the idea that the spectral-Poisson model is more accurate in predicting activities compared to the periodic Poisson process or random exploration.

10.3 Evaluation on Real Robot Exploration Tasks

As shown in the previous section, the Poisson-based exploration model with spectral methods for extracting periodic structure displays better performance than other types of exploration models in maximising observations of human activities. Let's also recall that, in the previous chapters, Bayesian inference for partially observable Poisson processes shows its benefit in improving the number of observations of human activities by correcting bias in the statistical estimates of the posterior distribution over human activities. Also, recall that the original intention of the thesis is to predict where many people are most likely to be and let an autonomous robot to go and observe them. Here, we put together those two different statistical models to predict where people are and drive exploration for human activities. This

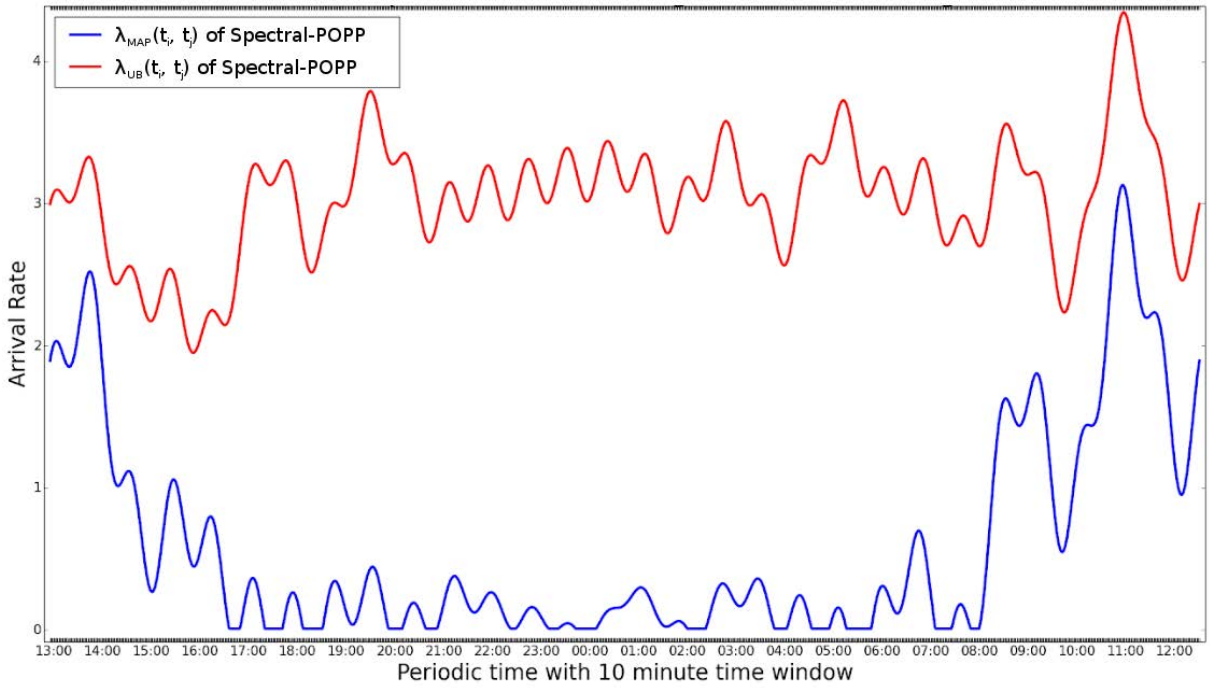

Figure 10.1: A spectral-Poisson process of region 9 represented by its MAP hypothesis ($\lambda_{MAP}(t_i, t_j)$) and its upper bound of the confidence interval $\lambda_{UB}(t_i, t_j)$.

section aims explicitly to show how the Bayesian inference for POPP together with spectral methods, as an exploration model, performs compared to the spectral-Poisson exploration model in regards to maximising observations of human activities.

Lets recall the dataset which had been used from Chapter 6 to Chapter 9 to assess the accuracy and performance of the POPP filters. This dataset[1] was, in fact, collected by the exploration models, as mentioned earlier, throughout 69-days of the robot deployment. Due to hardware failures, sensor malfunctions and other external issues related to the mobile robot, only 48 days from the dataset were deemed worth analysis.

The exploration models which were used during the deployment are the spectral-Poisson (or spectral-FOPP), the spectral-POPP, and the spectral-POPP-Beta. The spectral-FOPP follows the same procedure described in the previous section with one difference. Instead of having Δ set to a week period for the periodic Poisson process, the deployment set Δ to a day period.

The spectral-POPP and spectral-POPP-Beta are exploration models which follow a

[1]The dataset can be downloaded from https://github.com/ferdianjovan/spectral_popp

periodic Poisson model with Fourier transformation applied to the series of point estimates of the periodic model for getting a smoother model and POPP (or POPP-Beta) filter applied during the learning / updating process of the model for correcting any systematic error produced by detectors. In other words, these two models combine spectral-Poisson described in Chapter 4 with POPP models described in Chapter 6 (or 7 respectively). During robot's observation on some particular time interval t_i, t_j, the sensed counts were captured by the robots' detectors. These sensed counts were then used to do Bayesian update on $\lambda(t_i, t_j)$, i.e. $P_G(\lambda(t_i, t_j) \mid \alpha_{(i,j)}, \beta_{(i,j)})$, using Equation 6.2. The switching-filter described in Chapter 6 was chosen to approximate $P_G\big(\lambda(t_i, t_j) \mid \alpha_{(i,j)}, \beta_{(i,j)}\big)$ on each Bayesian update. The upper bound of the confidence interval ($\lambda_{UB}(t_i, t_j)$) of the updated arrival rate $\lambda(t_i, t_j)$ was selected for each $i, j \in \{1, \ldots, \Delta\}$ using Equation 10.1with Δ set to a day period for the periodic Poisson process. A series of upper bounds chosen for each $\lambda(t_i, t_j)$ were then encoded and extracted via spectral analysis with l-AAM technique to produce a smoother series of upper bound estimates $\lambda'_{UB}(t_i, t_j)$–following Algorithm 4. Figure 10.1 depicts a comparison between the MAP hypothesis estimate and the upper bound estimate of a spectral-Poisson process. Finally, places to visit throughout a day were chosen by Equation 10.2.

These three different exploration models were applied separately during the 69-day of the deployment. All of these models used the upper bound of the confidence interval of the arrival rate for their exploration policies. For the first 27 days of the deployment, the robot followed an exploration policy according to the spectral-FOPP model. This resulted in 18 days worth of data collected. From this 18 days data collected by following the spectral-FOPP exploration model, the last 3 days were used to train the sensor model needed for both the spectral-POPP and the spectral-POPP-Beta models. From day 28 to day 47, the robot followed an exploration policy according to the spectral-POPP model. This resulted in 15 days worth of data collected. Finally, from day 48 onwards, the robot followed an exploration policy according to the spectral-POPP-Beta model. This resulted to 15 days worth of data collected.

Similar to the evaluation on the Aruba dataset in the previous section, a comparison among different exploration strategies is made, based on how many observations the robot

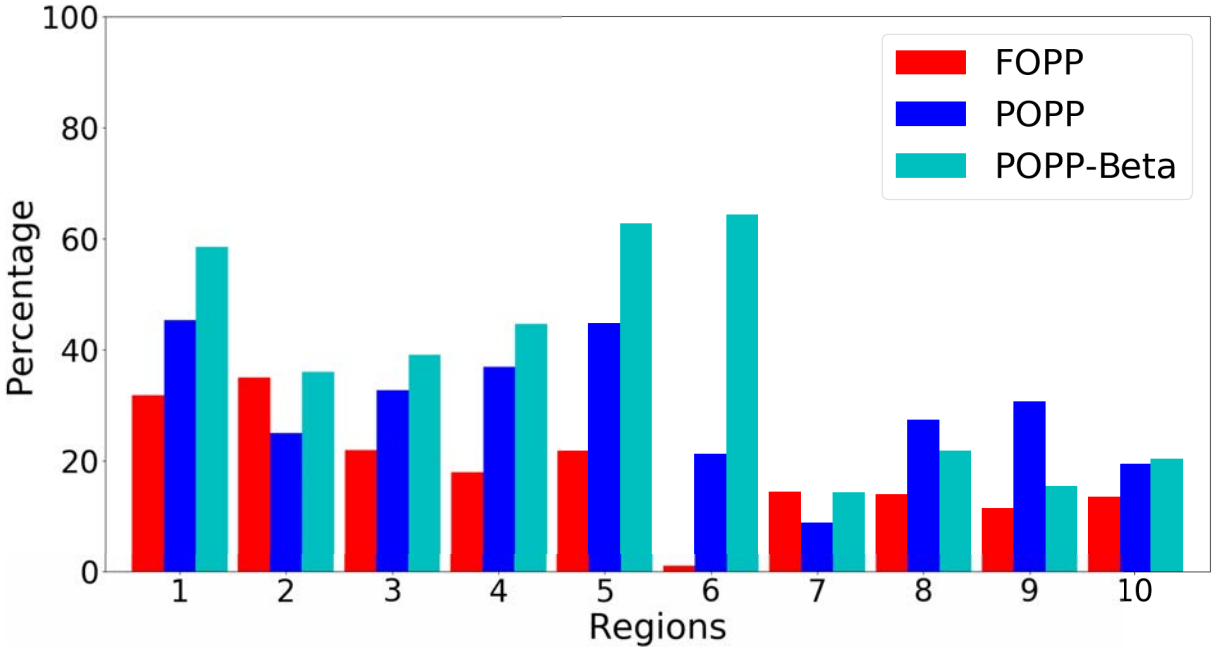

Figure 10.2: The activity exploration percentage across areas (regions) of the environment using three different exploration models (spectral-FOPP, spectral-POPP, spectral-POPP-Beta). The percentage shows the portion of time that the robot was observing activities.

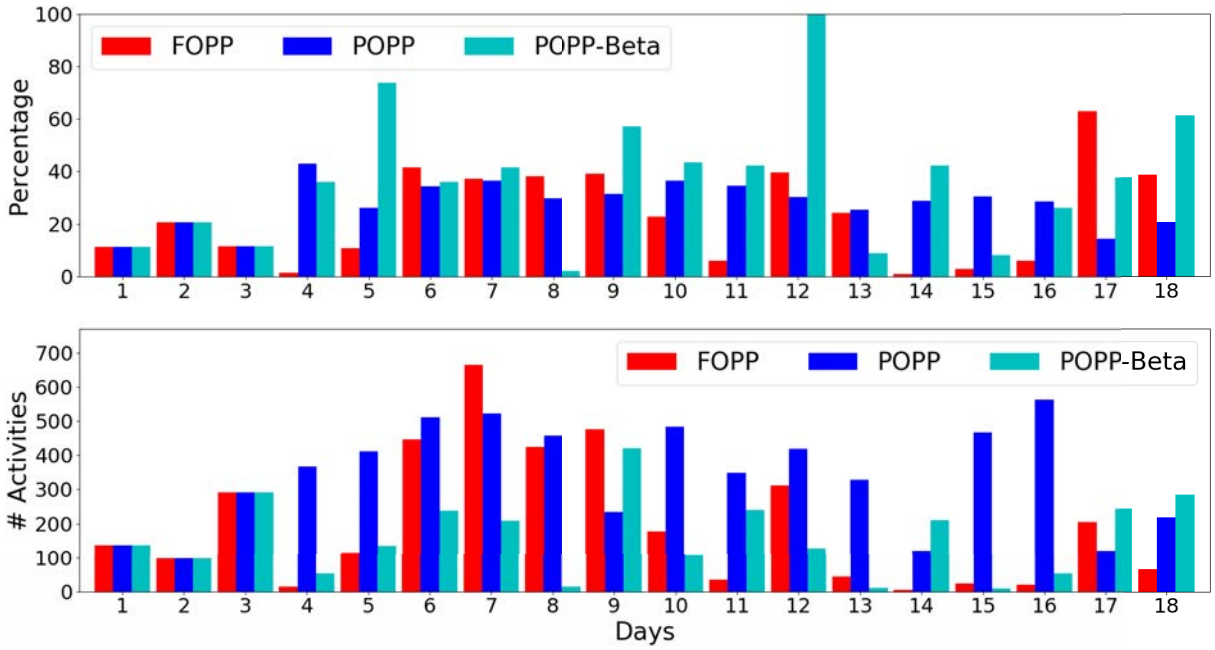

Figure 10.3: The exploration evolution from day 1 to day 15, averaged across all regions, for three exploration models. The top panel shows the portion of time that the robot was observing activities, whereas the bottom panel shows the number of activities.

makes using a particular exploration model. For this comparison, the last 3 days which are part of the 18 days worth of data collected by following the spectral-FOPP exploration model

are included in the spectral-POPP and the spectral-POPP-Beta exploration models. This is necessary to avoid the spectral-POPP and the spectral-POPP-Beta exploration models having an advantage over the spectral-FOPP exploration model since the spectral-POPP and the spectral-POPP-Beta need a training period to construct their sensor model. Moreover, due to the absence of information regarding activities happening in other places that the robot did not visit, only a comparison of the total duration of positive observations over the total duration of its visit can be compared and evaluated. As a reminder, a positive observation is a time when the robot observes any activity during its visit to a particular location.

Figure 10.2 shows how many positive observations the robot made in each region of the environment. As can be seen, the exploration policy produced by spectral-POPP-Beta has the highest positive observations in many of the regions followed by the exploration policy according to the spectral-POPP model. Recall that some regions, such as 4, 5, 6, and 7, are not densely populated with human activities across time compared to other regions (such as 1, 2, 3, and 10). The spectral-POPP and spectral-POPP-Beta models, however, still managed to improve the percentage of positive observations. This showed that the models correctly predicted that activities would take place in particular locations. One should note that region 6 is the place where there are vending machines. These are often detected as a person by one (upper body detector) of the detectors attached to the robot. This might lead the spectral-FOPP model to plan a visit to this particular location when no activity is taking place. As can be seen, the spectral-POPP and the spectral-POPP-Beta models were able to correct the miscounts occurring in region 6, and also giving a better estimate to the posterior over the arrival rate λ and thus making the exploration-exploitation trade-off using better beliefs.

A comparison of how the positive observations evolved with time for each exploration policies was recorded. This is depicted by Figure 10.3. It cannot necessarily be expected that the positive observations will increase monotonically as days pass. This is because different days will have a different structure of activities, and thus a higher or lower aggregate level of activity to detect. As shown in Figure 10.3, all exploration models showed similar performance

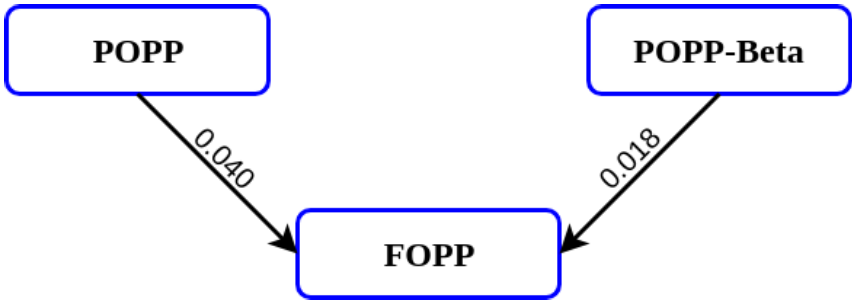

Figure 10.4: The significance relationship among the spectral-POPP, the spectral-POPP-Beta and the spectral-FOPP exploration models.

on the first three days because these are the last three days from the spectral-FOPP exploration model whose the data were used to train the sensor models of the spectral-POPP and spectral-POPP-Beta exploration models. The bottom panel of Figure 10.3 shows that the spectral-POPP-Beta exploration models did not observe many activities during the deployment, this is because the number of students that were around during the spectral-POPP-Beta exploration period were lower compared to other times when other exploration models (such as the spectral-POPP and the spectral-FOPP exploration model) were executed.

A test of significance was also performed to prove that results from the spectral-POPP, the spectral-POPP-Beta and the spectral-FOPP exploration models are caused by something other than randomness. The *p-value* = 0.05 was used as the cutoff for significance. If the *p-value* is less than 0.05, the null hypothesis that there is no difference between the means of the results is rejected, and a significant difference exists among the results of the spectral-POPP, the spectral-POPP-Beta, and the spectral-FOPP. The test of significance is shown in Figure 10.4. As shown in the figure, the results of the spectral-POPP and the spectral-POPP-Beta exploration models are significantly different than the one from the spectral-FOPP exploration model. However, the same conclusion can not be made between the result of the spectral-POPP and the result of the spectral-POPP-Beta as the *p-value* between those two results is above 0.05.

10.4 Discussion

This chapter discusses the application of the spectral-Poisson process with the POPP filter to control activity exploration performed by a mobile robot. Activity exploration was performed because the robot was only able to observe a limited portion of the lower-ground floor at any one time, and so it had to actively plan to go to other places to observe more activities. As has been shown, the exploration plan cannot be a random one. Any exploration model must optimise the exploration-exploitation trade-off to ensure good observation time of the robot.

The initial experiment on the Aruba dataset indicated that the exploration policy given by the spectral-Poisson model improved the optimisation of the exploration-exploitation trade-off over that of the periodic Poisson process. The policy, by extension, is better than the random exploration policy in efficiently finding activities.

The real experiment on a real robot exploring the lower-ground floor for a total of 69 days with the aim to observe more activities was extending the result shown on the Aruba dataset. By this experiment, the spectral-Poisson model for leveraging the periodic structures of human behaviour and the POPP model for correcting bias in statistical estimates were put together into one model, producing the spectral-POPP and the spectral-POPP-Beta models. This experiment not only created the UoB dataset but also provided support that the spectral POPP-Beta variant is the most effective method for optimising the exploration-exploitation trade-off.

Due to time limitations, the author could not evaluate the exploration policy according to both the N-POPP and POPP-Dirichlet model. One might expect that the N-POPP and the POPP-Dirichlet will improve the efficiency of the exploration policy further. However, to successfully implement the N-POPP and the POPP-Dirichlet model, one will need to have a larger initial dataset to train the joint sensor model.

CHAPTER 11

GENERAL DISCUSSION AND CONCLUSION

The problems addressed in thesis are motivated by an autonomous mobile robot which must adapt and learn dynamic behaviours from its unreliable perception to be able to work alongside humans in human-centered environments. The dynamic changes come from human activities which typically follow predictable, repeating patterns that generate corresponding changes in space. However, the robot's inability to see everywhere at once makes the learning process difficult due to incomplete information about its surroundings. Its perception algorithms, which are prone to systematic errors, introduce bias to its statistical inferences, aggravating the problem even further.

This thesis has been concerned with developing practical estimators for count data with regularities, collected by an autonomous mobile robot, with unreliable perception algorithms, on extended temporal scales. These count data represent the level of human activity in particular locations. It studies two problems: drawing inferences from 1) incomplete time series of count data and 2) unreliable count data. Several contributions have been made.

- A framework which allows us to sensibly retrieve and represent temporally recurring

patterns of count data on extended temporal scales from temporally sparse observations has been developed. The approach (Spectral-FOPP), based on a time-varying Poisson process in combination with spectral analysis, efficiently retrieves recurring patterns of count data. It is shown how to use these patterns to predict the aggregate human activity level at particular times and places. The framework has been extensively tested, both on simulated periodic count data and robot-gathered observations. The experiment also showed that the periodic patterns can be used to categorize locations. These categories are similar to the semantic categories given by humans.

- A set of inference methods for the partially observable Poisson process (POPP) has been formulated. The POPP is a Poisson process which takes into account the unreliability of the sensors that count events. Unlike Bayesian estimation for a fully observable Poisson process (FOPP), obtaining the posterior is non-trivial, since there is no conjugate density for a POPP and the posterior has a number of elements that grow exponentially in the number of observed intervals. Two simple, tractable, approximations have been presented. These two approximations are combined in a switching filter, which enables efficient and accurate estimation of the posterior. A simulation study shows that these POPP filters correct the over- and under-counts produced by sensors.

- Variations of the POPP filter are presented. The POPP-Beta filter extends the POPP filter in which the unreliability of the observation model is accounted for when estimations are built. The N-POPP filter extends the POPP filter by modelling the case when sensors are uncorrelated. The POPP-Dirichlet combines the POPP-Beta filter and the N-POPP filter to have the benefits of each correction. A simulation and observations taken by a robot, on a series of long deployments, show that each extension provides progressively more accurate estimates than the POPP filter.

- Both posteriors from the Spectral-FOPP and two Spectral-POPP processes are used to drive exploration by a mobile robot for a series of two week deployments. An upper bound interval exploration method was used to solve the exploration-exploitation

problem. After labelling by humans, this resulted in a labelled data set of six weeks of human activity levels. A simulated study has shown that the Spectral-FOPP and the Spectral-POPP filter improve on-point observation time significantly if strong periodic patterns underlying the human activities are present.

11.1 Limitations and Further Work

Two basic statistical models: Spectral-FOPP and POPP have been proposed and evaluated. The combination of these two is able to extract temporal dynamics in the aggregate level of human activities from unreliable sensors, along with the ability to exploit this understanding for better exploration by an autonomous mobile robot. However, the spectral-POPP model could still be improved in the following two ways:

1. In Chapter 6, The Gamma filter approximates a sum of Gamma distributions with a single Gamma distribution assuming that the sensor performs somewhat reliable. Instead of using a single Gamma distribution to approximate a sum of m Gamma distributions, n gamma distributions, where n is much smaller than m, could be used to improve the accuracy of the approximation to the posterior. This would promise to be more accurate than a single gamma, but more efficient than a histogram filter. Thus, it might be faster than the switching filter.

2. The spectral-Poisson model (Spectral-FOPP) in Chapter 4 is a statistical model which is able, and only able, to capture the periodic structure of count data. It indirectly assumes that there is an underlying pattern governing the evolution of the parameter λ of a Poisson process. The spectral-Poisson might not be able to capture other non-periodic structures governing the parameter λ, such as trends.

   A Gaussian process modulated Poisson process might provide a better model for different structures which govern λ over time. Work from Lloyd et al. (2015) presents a fully variational Bayesian inference scheme for continuous Gaussian-process modulated

Poisson process. It provides a good estimator and is fast in estimating λ of a Poisson process. An extension to this statistical model which embeds both trends and periodicity in the model might provide a solution to the limitations of Spectral-Poisson while being fully Bayesian.

REFERENCES

Aalen, O. (1978), 'Nonparametric inference for a family of counting processes', *The Annals of Statistics* pp. 701–726. 45

Achcar, J. (2001), 'Bayesian analysis for software reliability data', *Advances in Reliability* **20**(C), 733 – 748. 32

Achcar, J., Barrios, J. & Rodrigues, E. (2012), 'Comparing the adequacy of some non-homogeneous Poisson models to estimate ozone exceedances in mexico city', *Journal of Environmental Protection* **3**(9A), 1213–1227. 32

Alba, E. & Dorronsoro, B. (2005), 'The exploration/exploitation tradeoff in dynamic cellular genetic algorithms', *IEEE Transactions on Evolutionary Computation* **9**(2), 126–142. 2, 167

Aldrich, J. (1997), 'R. a. fisher and the making of maximum likelihood 1912-1922', *Statistical Science* **12**(3), 162–176. 11

Audibert, J.-Y., Munos, R. & SzepesvÃąri, C. (2009), 'ExplorationâĂŞexploitation tradeoff using variance estimates in multi-armed bandits', *Theoretical Computer Science* **410**(19), 1876 – 1902. Algorithmic Learning Theory. 2, 167

Bartlett, M. S. (1963), 'The spectral analysis of point processes', *Journal of the Royal Statistical Society. Series B (Methodological)* **25**(2), 264–296. 47

Bekele, B. N. (1998), Binomial data with misclassification: a Bayesian approach, PhD thesis, Baylor University. 72

Berger, J. O., Bernardo, J. M. & Sun, D. (2009), 'The formal definition of reference priors', *The Annals of Statistics* **37**(2), 905–938. 76

Bernardo, J. M. (2005), Reference analysis, *in* D. Dey & C. Rao, eds, 'Bayesian Thinking', Vol. 25 of *Handbook of Statistics*, Elsevier, pp. 17 – 90. 75, 76

Bityukov, S. & Krasnikov, N. (2000), 'On the observability of a signal above background', *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* **452**(3), 518 – 524. 75

Box, G. E., Jenkins, G. M., Reinsel, G. C. & Ljung, G. M. (2015), *Time series analysis: forecasting and control*, John Wiley & Sons. 39, 40

Bracewell, R. (1965), 'The fourier transform and iis applications', *New York* . 35

Brahim-Belhouari, S. & Bermak, A. (2004), 'Gaussian process for nonstationary time series prediction', *Computational Statistics and Data Analysis* **47**(4), 705 – 712. 43

Bratcher, T. L. & Stamey, J. D. (2002), 'Estimation of poisson rates with misclassified counts', *Biometrical Journal* **44**(8), 946–956. 73, 74

Brault, J. W. & White, O. R. (1971), 'The Analysis and Restoration of Astronomical Data via the Fast Fourier Transform', *Astronomy and Astrophysics* **13**, 169. 33, 34

Bross, I. (1954), 'Misclassification in 2 x 2 tables', *Biometrics* **10**(4), 478–486. 71

Buchanan, W. K., Hodges, P. & Theis, J. (2001), 'Which way the natural gas price: an attempt to predict the direction of natural gas spot price movements using trader positions', *Energy economics* **23**(3), 279–293. 39

Cameron, A. C. & Trivedi, P. K. (2013), *Regression analysis of count data*, Vol. 53, Cambridge university press. 9

CandÃĺs, E. J., Romberg, J. K. & Tao, T. (2006), 'Stable signal recovery from incomplete and inaccurate measurements', *Communications on Pure and Applied Mathematics* **59**(8), 1207–1223. 34

Carlin, B. P. & Louis, T. A. (2010), *Bayes and empirical Bayes methods for data analysis*, Chapman and Hall/CRC. 9, 33

Casadei, D. (2011), 'Statistical methods used in atlas for exclusion and discovery', *arXiv preprint arXiv:1108.2288* . 75

Casadei, D. (2012), 'Reference analysis of the signal + background model in counting experiments', *Journal of Instrumentation* **7**(01), P01012. 75, 76

Casadei, D. (2014), 'Reference analysis of the signal + background model in counting experiments ii. approximate reference prior', *Journal of Instrumentation* **9**(10), T10006. 77

Casadei, D., Grunwald, C., KrÃűninger, K. & Mentzel, F. (2017), 'Objective bayesian analysis of counting experiments with correlated sources of background', *Journal of Applied Statistics* **0**(0), 1–19. 75

Chen, T. T. (1979), 'Log-linear models for categorical data with misclassification and double sampling', *Journal of the American Statistical Association* **74**(366a), 481–488. 71, 72

Contreras, J., Espinola, R., Nogales, F. J. & Conejo, A. J. (2003), 'Arima models to predict next-day electricity prices', *IEEE Transactions on Power Systems* **18**(3), 1014–1020. 39

Cook, D. J. (2010), 'Learning setting-generalized activity models for smart spaces', *IEEE intelligent systems* **2010**(99), 1. 169

Coppola, C., Krajník, T., Duckett, T. & Bellotto, N. (2016), Learning temporal context for activity recognition, *in* 'ECAI'. 169

Coppola, C., Krajník, T., Duckett, T., Bellotto, N. et al. (2016), Learning temporal context for activity recognition., *in* 'European Conference on Artificial Intelligence', pp. 107–115. 33, 36, 52, 56

Cox, D. & Lewis, P. (1966), 'The statistical analysis of series of events'. 46

Cox, D. R. (1955), 'Some statistical methods connected with series of events', *Journal of the Royal Statistical Society. Series B (Methodological)* **17**(2), 129–164. 46, 47

Davison, A. C. & Ramesh, N. I. (1996), 'Some models for discretized series of events', *Journal of the American Statistical Association* **91**(434), 601–609. 30

DeGroot, M. H. & Schervish, M. J. (2012), *Probability and statistics*, Pearson Education. 9

Demortier, L., Jain, S. & Prosper, H. B. (2010), 'Reference priors for high energy physics', *arXiv:1002.1111v2* **82**, 034002. 75

Diggle, P. (1985), 'A kernel method for smoothing point process data', *Journal of the Royal Statistical Society. Series C (Applied Statistics)* **34**(2), 138–147. 46

Dondrup, C., Bellotto, N., Jovan, F., Hanheide, M. et al. (2015), Real-time multisensor people tracking for human-robot spatial interaction, *in* 'Proceedings of the 2015 International Conference on Robotics and Automation', ICRA/IEEE. 60, 81

Duckworth, P., Gatsoulis, Y., Jovan, F., Hawes, N., Hogg, D. C. & Cohn, A. G. (2016), Unsupervised learning of qualitative motion behaviours by a mobile robot, *in* 'Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems', AAMAS '16, International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, pp. 1043–1051. 60, 81

Duvenaud, D., Lloyd, J. R., Grosse, R., Tenenbaum, J. B. & Ghahramani, Z. (2013), Structure discovery in nonparametric regression through compositional kernel search, *in* 'Proceedings of the 30th International Conference on Machine Learning'. 43, 44, 58

Dvorzak, M. & Wagner, H. (2016), 'Sparse bayesian modelling of underreported count data', *Statistical Modelling* **16**(1), 24–46. 73

Feil-Seifer, D. & Matarìc, M. J. (2011), 'Socially assistive robotics', *IEEE Robotics Automation Magazine* **18**(1), 24–31. 1

Fentanes, J. P., Lacerda, B., Krajník, T., Hawes, N. & Hanheide, M. (2015), Now or later? predicting and maximising success of navigation actions from long-term experience, *in* '2015 IEEE International Conference on Robotics and Automation (ICRA)', IEEE, pp. 1112–1117. 36

Fink, D. (1997), A compendium of conjugate priors, PhD thesis, Montana State University. 9, 17, 53, 81

Ghassemi, N. H. & Deisenroth, M. P. (2014), Analytic long-term forecasting with periodic gaussian processes, *in* 'Proceeding of AISTATS', pp. 303–311. 42, 58

Girard, A., Rasmussen, C. E., Candela, J. Q. & Murray-Smith, R. (2003), Gaussian process priors with uncertain inputs application to multiple-step ahead time series forecasting, *in* 'Advances in neural information processing systems', pp. 545–552. 43

Grosse, R., Salakhutdinov, R., Freeman, W. & Tenenbaum, J. (2012), Exploiting compositionality to explore a large space of model structures, *in* 'Uncertainty in Artificial Intelligence'. 43

Guarnaccia, C., Quartieri, J., Barrios, J. M. & Rodrigues, E. R. (2014), 'Modeling environmental noise exceedances using non-homogeneous Poisson processes', *The Journal of the Acoustical Society of America* **136**(4), 1631–1639. 32, 33

Gutiérrez-Peña, E. & Nieto-Barajas, L. E. (2003), Bayesian nonparametric inference for mixed poisson processes, *in* 'Bayesian Statistics 7: Proceedings of the Seventh Valencia International Meeting', Oxford University Press, USA, p. 163. 48

Heikkinen, J. & Arjas, E. (2001), 'NonâĂŘparametric bayesian estimation of a spatial poisson intensity', *Scandinavian Journal of Statistics* **25**(3), 435–450. 47

Hochberg, Y. (1977), 'On the use of double sampling schemes in analyzing categorical data with misclassification errors', *Journal of the American Statistical Association* **72**(360a), 914–921. 71, 72

Hutchins, J., Ihler, A. & Smyth, P. (2007), Modeling count data from multiple sensors: A building occupancy model, *in* '2007 2nd IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing', pp. 241–244. 31

Hyndman, R. J. & Athanasopoulos, G. (2014), *Forecasting: principles and practice*, OTexts. 38, 39, 40

Ihler, A., Hutchins, J. & Smyth, P. (2006), Adaptive event detection with time-varying poisson processes, *in* 'Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining', ACM, pp. 207–216. 31, 80

Ihler, A., Hutchins, J. & Smyth, P. (2007), 'Learning to detect events with markov-modulated poisson processes', *ACM Transactions on Knowledge Discovery from Data* **1**(3). 31

Jovan, F., Wyatt, J. & Hawes, N. (2018), Efficient bayesian methods for counting processes in partially observable environments, *in* A. Storkey & F. Perez-Cruz, eds, 'Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics', Vol. 84 of *Proceedings of Machine Learning Research*, PMLR, Playa Blanca, Lanzarote, Canary Islands, pp. 1906–1913. 8

Jovan, F., Wyatt, J., Hawes, N. & Krajník, T. (2016), A Poisson-Spectral Model for Modelling the Spatio-Temporal Patterns in Human Data Observed by a Robot, *in* 'Proc IEEE Conf on Intelligent Robots and Systems', Daejeon, South Korea. 7, 8, 81

Karlin, S. (2014), *A first course in stochastic processes*, Academic press. 9

Karunaratne, P., Moshtaghi, M., Karunasekera, S., Harwood, A. & Cohn, T. (2017), Modelling the working week for multi-step forecasting using gaussian process regression, *in* 'Proceedings of the 26th International Joint Conference on Artificial Intelligence', AAAI Press, pp. 1994–2000. 42

Kingman, J. (2005), *Poisson Processes*, American Cancer Society. 9

Kirby, H. R., Watson, S. M. & Dougherty, M. S. (1997), 'Should we use neural networks or statistical models for short-term motorway traffic forecasting?', *International Journal of Forecasting* **13**(1), 43–50. 40

Kircher, T., Nelson, J. & Burdo, H. (1985), 'The autopsy as a measure of accuracy of the death certificate', *New England Journal of Medicine* **313**(20), 1263–1269. 73

Kottas, A. (2006), Dirichlet process mixtures of beta distributions, with applications to density and intensity estimation, *in* 'Workshop on Learning with Nonparametric Bayesian Methods, 23rd International Conference on Machine Learning (ICML)'. 47

Krajnik, T., Fentanes, J. P., Cielniak, G., Dondrup, C. & Duckett, T. (2014), Spectral analysis for long-term robotic mapping, *in* '2014 IEEE International Conference on Robotics and Automation (ICRA)', pp. 3706–3711. 33, 36, 49, 52, 56

Krajnik, T., Pulido Fentanes, J., Santos, J., Kusumam, K., Duckett, T. et al. (2015), Fremen: frequency map enhancement for long-term mobile robot autonomy in changing environments, *in* 'ICRA15 WS VPRiCE', IEEE. 36

Krajnık, T., Santos, J., Seemann, B. & Duckett, T. (2014), 'Froctomap: an efficient spatio-temporal environment representation', *Advances in Autonomous Robotics Systems* p. 269. 36

KrajnÃŋk, T., Fentanes, J. P., Mozos, O. M., Duckett, T., Ekekrantz, J. & Hanheide, M. (2014), Long-term topological localisation for service robots in dynamic environments using spectral maps, *in* '2014 IEEE/RSJ International Conference on Intelligent Robots and Systems', pp. 4537–4542. 33, 36, 37, 52, 56, 58

KrajnÃŋk, T., Kulich, M., MudrovÃą, L., Ambrus, R. & Duckett, T. (2015), Where's waldo at time t ? using spatio-temporal models for mobile robot search, *in* '2015 IEEE International Conference on Robotics and Automation (ICRA)', pp. 2140–2146. 33, 36, 52, 56

Kulis, B. & Jordan, M. I. (2012), Revisiting k-means: New algorithms via bayesian nonparametrics, *in* 'Proceedings of the 29th International Conference on Machine Learning (ICML)', pp. 513–520. 65

Kumar, U. (2015), 'An integrated ssa-arima approach to make multiple day ahead forecasts for the daily maximum ambient o3 concentration', *Aerosol and Air Quality Research* **15**(1), 208–219. 41

Kumar, U. & De Ridder, K. (2010), 'Garch modelling in association with fft–arima to forecast ozone episodes', *Atmospheric Environment* **44**(34), 4252–4265. 41

Laing, W. & Smith, D. (1987), A comparison of time series forecasting methods for predicting the cegb demand, *in* 'Proceedings of the ninth power systems computation conference'. 40

Leith, D. J., Heidl, M. & Ringwood, J. V. (2004), Gaussian process prior models for electrical load forecasting, *in* '2004 International Conference on Probabilistic Methods Applied to Power Systems', pp. 112–117. 43

Lloyd, C., Gunter, T., Osborne, M. & Roberts, S. (2015), Variational inference for gaussian process modulated poisson processes, *in* 'International Conference on Machine Learning', pp. 1814–1822. 46, 181

Lloyd, J. R. (2015), Representation, learning, description and criticism of probabilistic models with applications to networks, functions and relational data, PhD thesis, University of Cambridge. 43

Lloyd, J. R., Duvenaud, D., Grosse, R., Tenenbaum, J. B. & Ghahramani, Z. (2014), Automatic construction and Natural-Language description of nonparametric regression models, *in* 'Association for the Advancement of Artificial Intelligence (AAAI)'. 42, 43, 44

Lo, A. Y. (1982), 'Bayesian nonparametric statistical inference for poisson point processes', *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* **59**(1), 55–66. 45, 48

Lo, A. Y. & Weng, C.-S. (1989), 'On a class of bayesian nonparametric estimates: Ii. hazard rate estimates', *Annals of the Institute of Statistical Mathematics* **41**(2), 227–245. 45

Ludkovski, M. & Sezer, S. O. (2012), 'Finite horizon decision timing with partially observable poisson processes', *Stochastic Models* **28**(2), 207–247. 30, 80

Meier-Hellstern, K. (1987), 'A fitting algorithm for markov-modulated poisson processes having two arrival rates', *European Journal of Operational Research* **29**(3), 370 – 377. 30, 80

Moreno, E. & Girón, J. (1998), 'Estimating with incomplete count data a bayesian approach', *Journal of Statistical Planning and Inference* **66**(1), 147 – 159. 73

Mwalili, S. M. (2006), Bayesian and frequentist approaches to correct for misclassification error with applications to caries research., PhD thesis, Katholieke Universiteit Leuven. 5

MÃÿller, J., Syversveen, A. R. & Waagepetersen, R. P. (1998), 'Log gaussian cox processes', *Scandinavian Journal of Statistics* **25**(3), 451–482. 46

Nerlove, M. (1964), 'Spectral analysis of seasonal adjustment procedures', *Econometrica* **32**(3), 241–286. 33

Nochai, R. & Nochai, T. (2006), Arima model for forecasting oil palm price, *in* 'Proceedings of the 2nd IMT-GT Regional Conference on Mathematics, Statistics and Applications', pp. 13–15. 39

Papoulis, A. & Chamzas, C. (1979), 'Detection of hidden periodicities by adaptive extrapolation', *IEEE Transactions on Acoustics, Speech, and Signal Processing* **27**(5), 492–500. 33

Papoulis, A. & Pillai, S. U. (2002), *Probability, random variables, and stochastic processes*, Tata McGraw-Hill Education. 9, 84

Peeling, P., fai Li, C. & Godsill, S. (2007), 'Poisson point process modeling for polyphonic music transcription', *The Journal of the Acoustical Society of America* **121**(4), EL168–EL175. 46

Pierini, M., Prosper, H. B., Sekmen, S. & Spiropulu, M. (2011), 'Priors for new physics', *arXiv preprint arXiv:1108.0523* . 75

Prabhu, N. U. & Zhu, Y. (1989), 'Markov-modulated queueing systems', *Queueing Systems* **5**(1), 215–245. 30, 80

Rasmussen, C. E. (2004), Gaussian processes in machine learning, *in* 'Advanced lectures on machine learning', Springer, pp. 63–71. 41, 42

Rasmussen, C. E. & Ghahramani, Z. (2001), Occam's razor, *in* 'Advances in neural information processing systems', pp. 294–300. 45

Rathbun, S. L. & Cressie, N. (1994), 'Asymptotic properties of estimators for the parameters of spatial inhomogeneous poisson point processes', *Advances in Applied Probability* **26**(1), 122–154. 46

Rue, H., Martino, S. & Chopin, N. (2009), 'Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **71**(2), 319–392. 48

RydÃl'n, T. (1996), 'An em algorithm for estimation in markov-modulated poisson processes', *Computational Statistics and Data Analysis* **21**(4), 431 – 447. 30, 80

Satterthwaite, F. E. (1946), 'An approximate distribution of estimates of variance components', *Biometrics bulletin* **2**(6), 110–114. 88

Scargle, J. D. (1982), 'Studies in astronomical time series analysis. II - Statistical aspects of spectral analysis of unevenly spaced data', *Astrophysical Journal* **263**, 835–853. 33, 34

Schaidnagel, M. & Laux, F. (2013), 'Time series prediction with automated periodicity detection', *International Journal on Advances in Systems and Measurements* **6**(3 & 4), 394–404. 39

Schmid, G., Goychuk, I. & Hänggi, P. (2006), 'Capacitance fluctuations causing channel noise reduction in stochastic hodgkinâĂŞhuxley systems', *Physical Biology* **3**(4), 248. 81

Scott, S. L. (1998), Bayesian methods and extensions for the two state Markov modulated Poisson process, PhD thesis, Harvard University Cambridge, Massachusetts. xi, 9, 30, 31, 80

Scott, S. L. (2001), 'Detecting network intrusion using a markov modulated nonhomogeneous poisson process', *Submitted to the Journal of the American Statistical Association* . 31, 80

Scott, S. L. & Smyth, P. (2003), 'The markov modulated poisson process and markov poisson cascade with applications to web traffic data', *Bayesian statistics* **7**, 671–680. 31

Shu, Y., Yu, M., Liu, J. & Yang, O. W. W. (2003), Wireless traffic modeling and prediction using seasonal arima models, *in* 'Communications, 2003. ICC '03. IEEE International Conference on', Vol. 3, pp. 1675–1679 vol.3. 40

Sposto, R., Preston, D. L., Shimizu, Y. & Mabuchi, K. (1992), 'The effect of diagnostic misclassification on non-cancer and cancer mortality dose response in a-bomb survivors', *Biometrics* **48**(2), 605–617. 73

Stamey, J. D. & Young, D. M. (2005), 'Maximum likelihood estimation for a poisson rate parameter with misclassified counts', *Australian & New Zealand Journal of Statistics* **47**(2), 163–172. 74

Stamey, J. D., Young, D. M. & Bratcher, T. L. (2004), 'Bayesian predictive probability functions for count data that are subject to misclassification', *Biometrical Journal* **46**(5), 572–578. 74

Stark, H. (1987), *Image recovery: theory and application*, Elsevier. 34

Taylor, J. W. (2003), 'Short-term electricity demand forecasting using double seasonal exponential smoothing', *Journal of the Operational Research Society* **54**(8), 799–805. 40

Templeton, M. (2004), 'Time-series analysis of variable star data', *Journal of the American Association of Variable Star Observers* **32**, 41–54. 33, 35, 56

Tenenbein, A. (1970), 'A double sampling scheme for estimating from binomial data with misclassifications', *Journal of the American Statistical Association* **65**(331), 1350–1361. 71, 72

Tijms, H. C. (2003), *A first course in stochastic models*, John Wiley and sons. 9

Tseng, F.-M., Yu, H.-C. & Tzeng, G.-H. (2002), 'Combining neural network model with seasonal time series arima model', *Technological Forecasting and Social Change* **69**(1), 71 – 87. 40

Tularam, G. A. & Ilahee, M. (2010), 'Time series analysis of rainfall and temperature interactions in coastal catchments', *Journal of Mathematics and Statistics* **6**(3), 372–380. 41

Turin, W. (1996), 'Fitting probabilistic automata via the em algorithm', *Stochastic Models* **12**(3), 405–424. 31

Varuzza, L., Gruber, A. & Pereira, C. A. (2008), 'Significance tests for comparing digital gene expression profiles', *arXiv preprint arXiv:0806.3274* . 5

Viana, M., Ramakrishnan, V. & Levy, P. (1993), 'Bayesian analysis of prevalence from the results of small screening samples', *Communications in Statistics - Theory and Methods* **22**(2), 575–585. 71, 72

Waagepetersen, R. (2004), 'Convergence of posteriors for discretized log gaussian cox processes', *Statistics and Probability Letters* **66**(3), 229 – 235. 46, 47, 48

Waagepetersen, R. P. (2007), 'An estimating function approach to inference for inhomogeneous neymanâĂŞscott processes', *Biometrics* **63**(1), 252–258. 48

Weiss, E. (2000), 'Forecasting commodity prices using arima', *Technical Analysis of Stocks and Commodities* **18**(1), 18 – 19. 39

Welch, B. L. (1947), 'The generalization ofstudent's' problem when several different population variances are involved', *Biometrika* **34**(1/2), 28–35. 88

Whittemore, A. S. & Gong, G. (1991), 'Poisson regression with misclassified counts: Application to cervical cancer mortality rates', *Journal of the Royal Statistical Society. Series C (Applied Statistics)* **40**(1), 81–93. 72

Williams, B. M. & Hoel, L. A. (2003), 'Modeling and forecasting vehicular traffic flow as a seasonal arima process: Theoretical basis and empirical results', *Journal of transportation engineering* **129**(6), 664–672. 39

Williams, C. K. & Rasmussen, C. E. (2006), 'Gaussian processes for machine learning', *the MIT Press* **2**(3), 4. 43

Wilson, A. & Adams, R. (2013), Gaussian process kernels for pattern discovery and extrapolation, *in* 'Proceedings of the 30th International Conference on Machine Learning (ICML-13)', pp. 1067–1075. 43

Wolpert, R. L. & Ickstadt, K. (1998), 'Poisson/gamma random field models for spatial statistics', *Biometrika* **85**(2), 251–267. 47

Wyatt, J. (1998), Exploration and inference in learning from reinforcement, PhD thesis, University of Edinburgh. College of Science and Engineering. School of Informatics. 2, 167