Master's thesis

NTNU
Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Department of Mathematical Sciences

Simon Liabø

Modeling Passenger Count Data Based on Automatic Counting

Master's thesis in Natural Science with Teacher Education
Supervisor: Ingelin Steinsland

June 2023

NTNU
Norwegian University of
Science and Technology

Simon Liabø

Modeling Passenger Count Data Based on Automatic Counting

Master's thesis in Natural Science with Teacher Education
Supervisor: Ingelin Steinsland
June 2023

Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Department of Mathematical Sciences

NTNU
Norwegian University of
Science and Technology

Abstract

Equipped with a host of sensors, modern public transport fleets generate a wealth of data, like Automatic Vehicle Location (AVL) and Automatic Passenger Counts (APC), although errors in this automatically collected data presents a challenge. The collaborative APT-R project seeks to realize the potential of automatic sensor data from public transport vehicles, through the development of innovative methods and tools.

This thesis considers APC data supplied by AtB, the public transport operator in Trondheim, Norway. The aim is to better understand how the APC data reflects the true passenger count (PC), and the creation of models to better utilize this data. We develop probabilistic models for the true PC of boarding passengers at door-level, using APC data. For this purpose, two modeling approaches were employed.

We first used the framework of generalized linear models (GLMs), for PC with APC as the explanatory variable. For the response, PC, we first considered the Poisson distribution, and then the double Poisson to account for the significant underdispersion. The double Poisson was inadequate in accounting for all the underdispersion, due to peaks in the data where APC is correct in its count of PC. Therefore, we make use of the k-inflated double Poisson distribution, and propose an extension where the inflation point k is allowed to vary with the discrete, explanatory variable APC. k-inflated distributions have been employed in regression settings previously, but only with fixed inflation points.

Further, a model that is based in the data generating process is proposed. The data generating process is considered the combination of an undercounting- and an overcounting process, modeled as a Binomial and a Poisson respectively. Empirical Bayes is used to do inference on the true passenger count for this model.

The models are fitted to, and evaluated on, door-level counts. In addition we investigate their performance on stop- and journey-level aggregates.

Our extension of the k-inflated double Poisson GLM shows promising results, improving on the double Poisson and demonstrating good fit to door-level counts. It also outperforms the model based in the data generating process, though this model also provides a good fit considering the relatively strict assumptions. Investigation of stop- and journey-level aggregates show clear dependency between door-level observations within these groups. This results in non-satisfactory uncertainty quantification for aggregates.

Sammendrag

Moderne offentlige transportkjøretøy er utstyrt med en rekke sensorer og genererer store mengder data, som Automatisk Kjøretøy Lokalisering (Automatic Vehicle Location, AVL) og Automatisk Passasjertelling (Automatic Passenger Counts, APC). Feil i disse dataene er en utfordring, og samarbeisprosjektet APT-R har som mål å realisere potensialet i automatisk sensordata fra kollektivsystemet, gjennom utvikling av innovative metoder og verktøy.

Denne masteroppgaven tar for seg APC-data levert av AtB, kollektivtransportselskapet i Trondheim. Målet er å bedre forstå hvordan APC-dataene gjenspeiler den virkelige passasjertellingen (PC), og å lage modeller for å bedre ta i bruk denne dataen. Vi utvikler probabilistiske modeller for den sanne tellingen av påstigende passasjerer på dørnivå, ved bruk av APC-data. Til dette formålet ble to modelleringsmetoder tatt i bruk.

Vi brukte først det fleksible rammeverket for generaliserte lineære modeller (GLM), med PC som respons og APC som forklarende variabel. For responsen, PC, vurderte vi først Poisson-fordelingen, og deretter double-Poisson-fordelingen for å ta hensyn til den betydelige underspredningen (underdispersion). Double-Poisson-fordelingen klarte ikke å ta hensyn til all underspredningen, dette skyldtes topper i dataene hvor APC er korrekt. Vi bruker derfor k-inflated double-Poisson-fordelingen, og foreslår en utvidelse av denne hvor inflasjonspunktet k får variere med den diskrete, forklaringsvariabelen APC. k-inflated-fordelinger har vært brukt i regresjonssammenhenger tidligere, men kun med fastsatte inflasjonspunkter.

Videre foreslår vi en modell som er basert på den data-genererende prosessen. Den data-genererende prosessen antas å være kombinasjonen av en undertellingsprosess og en overtellingsprosess, som vi modellerer med henholdsvis en binomial- og en Poisson-fordeling. For denne modellen tar vi i bruk empirisk Bayes metode for å gjøre inferens på den sanne passasjertellingen.

Modellene er tilpasset, og evaluert på, tellinger på dørnivå. I tillegg undersøker vi modellprestasjon på tellinger som er aggregert til stopp- og turnivå.

Vi ser lovende resultater for den implementerte k-inflated double Poisson GLM-en. Den er en forbedring av double-Poisson-fordelingen og viser generelt god tilpasning til tellinger på dørnivå. Den yter også bedre enn modellen som baserer seg på den data genererende prosessen, selv om denne modellen også viser god tilpasning, til tross for de relativt strenge antagelsene. Undersøkelse av aggregater på stopp- og reisenivå viser tydelig avhengighet mellom tellinger på dørnivå, noe som resulterer i dårlig usikkerhetskvantifisering av aggregater.

Foreword

This thesis concludes my Master's degree in Natural Science With Teacher Education, undertaken at the Norwegian University of Science and Technology (NTNU). My specialization in statistics enriches my capabilities as an educator. It provides me with a depth of knowledge from which I can draw, to foster a comprehensive understanding of the subject matter among my students.

I would like to express my gratitude to my supervisor, Ingelin Steinsland, whose invaluable guidance and discussions have been of great help. Further, a note of thanks also goes to the people at AtB for their cooperation, and for being a joy to work with.

A special thanks goes to Studentersamfundet, and the great people I met there. You have made my experience as a student in Trondheim a truly awesome and memorable chapter of my life.

Lastly, my deepest gratitudes are reserved for family and friends.

Table of Contents

List of Figures

List of Tables

Abbreviations

Abbreviation	*Description*
AIC	Akaike Information Criterion
APC	Automatic Passenger Count
CDF	Cumulative Density Function
(C)RPS	(Continuous) Ranked Probability Score
DGP	Data Generating Process
EDA	Exploratory Data Analysis
GLM	Generalized Linear Model
KIDPO	k-inflated double Poisson
MLE	Maximum Likelihood Estimate
PC	Passenger Count
PIG	Poisson-Inverse Gaussian
PIT	Probability Integral Transform
PMF	Probability Mass Function

1

Introduction

Recent societal trends highlight the need for sustainable transportation systems, and data from modern public transport fleets, are instrumental in this quest. The wealth of archived and real-time automatic sensor data, particularly Automatic Vehicle Location (AVL) and Automatic Passenger Counting (APC), harbors untapped potential for enabling data-driven monitoring, planning, and execution of sustainable, efficient public transport services.

The APT-R project is a joint effort among the Institute of Transport Economics (Transportøkonomisk institutt), the Norwegian University of Science and Technology (NTNU), the Norwegian Computing Center, and public transport operators Entur, Kolumbus, and AtB. This initiative strives to create innovative methods and tools to extract valuable insights from public transport automatic sensor data (Transportøkonomisk institutt, 2022). In this thesis, we examine automatic passenger counting (APC) data collected from buses operating in the Norwegian town of Trondheim. We aim to develop probabilistic models that predict actual passenger count (PC) using APC data, in order to better understand the relationship between these two variables.

APC is a technology developed to automatically count the number of passengers boarding and alighting public transport vehicles by utilizing various sensors. These sensors detect the entry and exit of passengers, thereby equipping public transport operators with valuable information to enhance management and optimization of their services. Over the past few decades, the deployment of APC technology has witnessed substantial growth, fueled by consistent advancements in sensor technology and the increasingly sophisticated techniques employed in data processing.

Mccarthy et al. (2021) gives a good account of the most widely adopted APC technologies, namely floor-based, WiFi, infrared and video sensing systems. Among these, video-based counting, leveraging algorithms rooted in pre-trained convolutional neural networks is the most commonly used (Mccarthy et al., 2021). This technology is also employed by AtB. Mccarthy et al. (2021) further notes that while video-based systems offer a wealth of information, they also pose considerable challenges that must be addressed by the software. These include distinguishing between people and objects, handling individuals moving in close proximity, and dealing with passengers already on board appearing in the frame. For the purposes of this thesis this implies a complex data generating process (DGP).

AtB (2016) state that their purpose in counting passengers is to collect data on board-

ing and alighting passengers and compile this data into statistics. These statistics can provide insights into passenger load on routes and journeys, which in turn is valuable information for route planning and reporting. AtB also shares passenger statistics with local and national authorities, serving as essential input for informed decision-making, and notes the importance of reliable passenger statistics for such purposes (AtB, 2016).

The increasingly enormous amount of available APC data, presents opportunities beyond the reporting of statistics. Nagaraj et al. (2022) proposes the application of deep learning for short-term passenger flow models using APC data. Similarly, Halyal et al. (2022) use APC data to forecast passenger demand using neural networks. Mccarthy et al. (2021) highlight the potential of APC in providing real-time passenger load estimates for planning and customer information.

Berrebi et al. (2022) notes that despite the vast amount of information available to public transport agencies through APC data, it has rarely been used to its' full potential due to concerns about the data quality. Much of the prior research has primarily focused on evaluating the accuracy of APC data by cross-referencing it with manual counts, counts from fare collection, and counts from video sources (Boyle (2008), as cited in Berrebi et al. (2022)). Strathman (1989) cross-checks with manual counts and Kimpel et al. (2003) with counts from video sources, both finding APC to be consistent with the respective reference data (Berrebi et al., 2022, p.2).

AtB, on the other hand, recognizes that their present APC system has a tendency to systematically underestimate passenger data. In response they apply a scalar adjustment to the APC data when reporting passenger counts. Although this method may be adequate when handling large volumes of data, its effectiveness for smaller subsets of selected data remains uncertain. Moreover, this approach does not quantify the uncertainty associated with the reported figures, and is not applicable in real-time scenarios.

Developing a model that establishes a connection between PC and APC data would give a deeper understanding of how APC data reflects PC. By analyzing this relationship, such a model could provide valuable context and insights when utilizing APC data, allowing for more accurate assessments and informed decision making. Ultimately, this approach has the potential to improve the reliability of reported passenger count data and contribute to the effective management of public transport services.

It is evident that the true DGP operates from PC to APC, and APC can be considered an error-prone meassurement of the true, underlying variable PC. Cameron and Trivedi (2013) note that the classical measurement error model with normal errors is inappropriate for count variables as it violates their non-negativity and discreteness. They instead introduce some alternative parametric models, each necessitating distinct assumptions on the mechanisms of the measurement error. In order to preserve non-negativity, the error model must necessarily be a finite mixture to account for both over- and underestimation. The selection of an appropriate finite mixture error model will require a thorough understanding of the complex DGP. Additionally, we will need a prior on PC to obtain a model for PC conditional on APC, and this approach rarely produces a parametric model for PC.

Alternatively it would be convenient to avoid making assumptions on the DGP, and employ a pure statistical approach. A generalized linear model (GLM) for PC with APC as the explanatory variable, lets us model a linear relation from the readily available APC to the variable of interest, PC. This yields a parametric probabilistic function for PC which could be easily employed by public transport operators in a variety of

scenarios.

In addition to the GLM, we also construct a model aiming to capture the true DGP. APC is modeled as a finite mixture, with PC as a parameter. Utilizing Bayes' theorem with an empirical prior on PC, we then derive a model for PC conditional on APC.

Both approaches results in models for PC with quantified uncertainty, which we call probabilistic models. The developed probabilistic models for PC use door-level observations of APC. We also explore how the models perform when applied to stop- and journey-level aggregates, offering a more comprehensive understanding of the behavior and usefulness of these models in various scenarios.

The text is structured as follows:

Chapter 2 outlines our data's origin, collection process, observed errors and implications, concluding with an initial exploratory data analysis that guides the subsequent modeling procedure.

Chapter 3 outlines the theoretical foundations of the statistical methods and models applied in Chapter 4. The focus is primarily on count data, its attributes, probability functions, and linear count regression. It further explores maximum likelihood estimation and Bayesian inference, concluding with a discussion on model assessment criteria, including AIC, (C)RPS and PIT histograms.

Chapter 4 proposes two door-level probabilistic models for PC conditional on APC: a GLM with APC as an explanatory variable using Poisson, double Poisson, and k-inflated double Poisson distributions, and DGP-M, which emulates the true data generating process. After detailing the models and their application to stop- and journey-level aggregates, the chapter ends with a section on model evaluation and an overview of the utilized software.

Chapter 5 presents results for the probabilistic models for PC as detailed in Chapter 4. These include parameter estimates, fitted PMFs, AIC- and (C)RPS-values, and PIT histograms. We first assess four door-level models for PC, three GLMs and the DGP-M model, then proceed to the results for stop- and journey-level aggregates.

Chapter 6 gives a discussion of the results, and provides some final remarks.

Appendix A is a derivation of expressions for the approximate expectation and variance of the k-inflated double Poisson distribution. These were ultimately found to deviate too much from the numerical calculations.

2

Data

In this chapter, we detail the specifics of the data used in the thesis. We start by discussing the origin of the data, and the data collection process. Then, we make some observations and discuss sources of errors discovered during data collection, and the further implications of these. The chapter concludes with some exploratory data analysis (EDA), providing some initial insights for the modeling procedure.

2.1 Data Collection and Preprocessing

This thesis utilizes data provided by AtB, the public bus operator in Trondheim, Norway. The dataset includes passenger count (PC) of passengers boarding and alighting the bus through each door during a stop, as well as the corresponding automatic passenger count (APC) for both boarding and alighting passengers. All of AtB's buses in Trondheim come equipped with DILAX optical sensors that utilize 3D stereo vision technology (DILAX). These sensors are also capable of providing video recordings. By capturing video from the sensors and manually counting passengers, we have gathered data on the actual number of passengers (PC). Though some errors may have occurred during manual counting, the PC data collected is considered to be the ground truth.

The data set comprises data from five selected bus routes: 1, 3, 10, 11, and 14. These routes operate on some of the same stops and are serviced by three different models of buses. Route 1 is a popular commuter route, its journey between the residential areas Heimdal and Ranheim taking it through the city center. Route 3 is also used by commuters as well as students, journeying through the residential Byåsen, the city center, and university campuses at Gløshaugen and Dragvoll. Routes 1 and 3 often carry a high volume of passengers throughout their journeys, and are operated by high capacity metro buses, $bus\_type = "metro"$. For these routes, high passenger counts are not uncommon, especially at the central "Kongens gate" and "Prinsens gate" bus stops, as well as the bus stops at university campuses.

The routes 10, 11 and 14 are operated by lower capacity buses, $bus\_type = "ordinary"$, and make more sprawling journeys along less trafficked roads. We expect lower passenger counts for these routes, but 10 and 11 follow stretches of the heavily trafficked "Elgeseter-/Prinsens gate" and will see an uptick in this area.

The selection of data provides a cross-section of highly populated and lower populated routes, as well as a variety of bus models, for a somewhat overlapping set of bus

stops. The data was collected in the period December 2022, through January 2023. It is important to note that due to the manual nature of video recording, all data has been collected during weekdays and within working hours (7:00-17:00).

For the data collection procedure we used screen recordings of the live video feed from the APC sensors, and of the real-time APC recordings. AtB uses backend logic that groups the boarding and alighting APC from all four doors and matches this stop-level count with a bus stop. This logic sometimes places an APC observation at the wrong stop, and the passengers boarding or alighting the bus at the last stop of a journey might be registered for the first stop of the next journey. When manually counting passengers from the screen recordings we were able to correct for these errors placing both PC and APC at the correct bus stop and door. Thus we can be sure that every paired PC and APC correspond to the same counting instance.

Another substantial source of error were discovered during data collection. In AtB's implementation the sensor is activated when the corresponding door is registered as open. The door registration is sometimes slow, resulting in several passengers being able to board and/or alight before the sensor is activated. This flaw in the system can thus create extreme outliers. It has by far the biggest impact on alighting data, as alighting passengers tend to go first and are often ready to leave the bus the instant the door opens. As many as 10 passengers can sometimes alight before the sensor is activated. Boarding data is also affected, but not nearly as much. AtB is now aware of the flaw and are working on adapting their APC system. In the remaining chapters we will refer to this source of error as "door-issues".

Given that this dominant source of error primarily affects alighting data, we have decided to focus exclusively on boarding passengers for this thesis. A model solely for boardings can still adequately report on the total number of passengers, as each passenger is accounted for upon entry, but a model for alightings will be needed for applications such as load estimation. It is further our opinion that insights gained for boarding passengers should be useful for alighting passengers as well.

2.2 Data Description

The data set comprises 2752 door-level counts, and the available variables are presented in Table 2.1. Our key variables of focus are PC and the corresponding APC. In addition the variables $door$, bus_stop, $stop$, $route$ and $journey$ are available, and give context to the counts. The true and automatic passenger counts are represented by the variables PC and APC respectively. bus_stop denotes the name of the bus stop where the passenger counts are recorded. $stop$ is a categorical factor that signifies one instance of a bus stopping at a bus stop, it encompasses four distinct counts, one for each of the bus's four doors. The $door$ variable is another factor that specifies at which of these four doors the counts were logged. $route$ indicates the specific route or pathway that the bus follows, and bus_type whether it is operated by a "metro" bus or an "ordinary" bus. Finally, $journey$ is a categorical factor that signifies one instance of a bus traveling along one route, it consolidates all counts from the start to the end of a single journey or trip. Table 2.2 gives an illustration of the data structure.

PC	APC	$door$	$stop$	$journey$	$route$	bus_stop	bus_type
0	0	1	1.159	1	”3”	”Skansen”	”metro”
3	4	2	1.159	1	”3”	”Skansen”	”metro”
5	3	3	1.159	1	”3”	”Skansen”	”metro”
0	0	4	1.159	1	”3”	”Skansen”	”metro”
0	1	1	1.160	1	”3”	”Ila”	”metro”
2	2	2	1.160	1	”3”	”Ila”	”metro”
4	0	3	1.160	1	”3”	”Ila”	”metro”
0	2	4	1.160	1	”3”	”Ila”	”metro”

Table 2.2: Example of data collected during two stops of one journey. (Not actual data, only for illustration purposes.)

Variable	Type	Description
PC	Integer	The true count of passengers boarding the bus.
APC	Integer	The corresponding automatic count.
$door$	Factor	The id of a door. 1 through 4, front to back.
bus_stop	String	The name of the bus stop.
$stop$	Factor	The unique id of a stop.
$route$	String	The name of the route.
$journey$	Factor	The unique id of a journey.
bus_type	String	”metro” or ”ordinary”

Table 2.1: Description of variables.

2.3 Exploratory Data Analysis (EDA)

2.3.1 Door-Level

The data set has values for APC in the range $[0, 14]$, and PC in the range $[0, 16]$. Tables 2.3 and 2.4, shows how the number of data points are distributed across these ranges, along with the accuracy of APC. We see that $APC = 0$ constitutes about 71% of the data set. APC is also highly accurate in this case, when $APC = 0$ it is correct 96.44% of the time. Consequently, a model for PC that predicts $PC = 0$ all the time would perform fairly well and we should consider fitting $APC = 0$ separately from the rest of the data. We also note that APC has an accuracy of 99.63% when $PC = 0$, in fact APC only registers a value other than 0, 7 out of the 1904 instances of $PC = 0$.

Logically, the larger counts represent a relatively larger fraction of total passengers compared to their proportion of data points, and a useful model would need perform adequately on the entire range of APC. However, as APC increases, the number of data points dramatically decreases. We also note that as APC increases, its accuracy seem to decrease, though not necessarily linearly. The goal should therefore be to apply

APC	n	%-of-data	accurate (PC=APC)	%-accuracy
0	1967	71.48	1897	96.44
1	368	13.37	297	80.71
2	192	6.98	139	72.40
3	100	3.63	63	63.00
4	44	1.60	24	54.55
5	26	0.94	15	57.69
6	20	0.73	13	65.00
7	14	0.51	8	57.14
8	8	0.29	2	25.00
9	7	0.25	2	28.57
10	3	0.11	1	33.33
11	1	0.04	0	0.00
13	1	0.04	0	0.00
14	1	0.04	0	0.00

Table 2.3: Distribution of data points on the range of APC, with accuracy.

PC	n	%-of-data	accurate (APC=PC)	%-accuracy
0	1904	69.19	1897	99.63
1	366	13.3	297	81.15
2	211	7.67	139	65.88
3	117	4.25	63	53.85
4	50	1.82	24	48
5	36	1.31	15	41.67
6	21	0.76	13	61.9
7	16	0.58	8	50
8	10	0.36	2	20
9	7	0.25	2	28.57
10	6	0.22	1	16.67
12	2	0.07	0	0.00
13	2	0.07	0	0.00
14	2	0.07	0	0.00
16	2	0.07	0	0.00

Table 2.4: Distribution of data points on the range of PC, with accuracy.

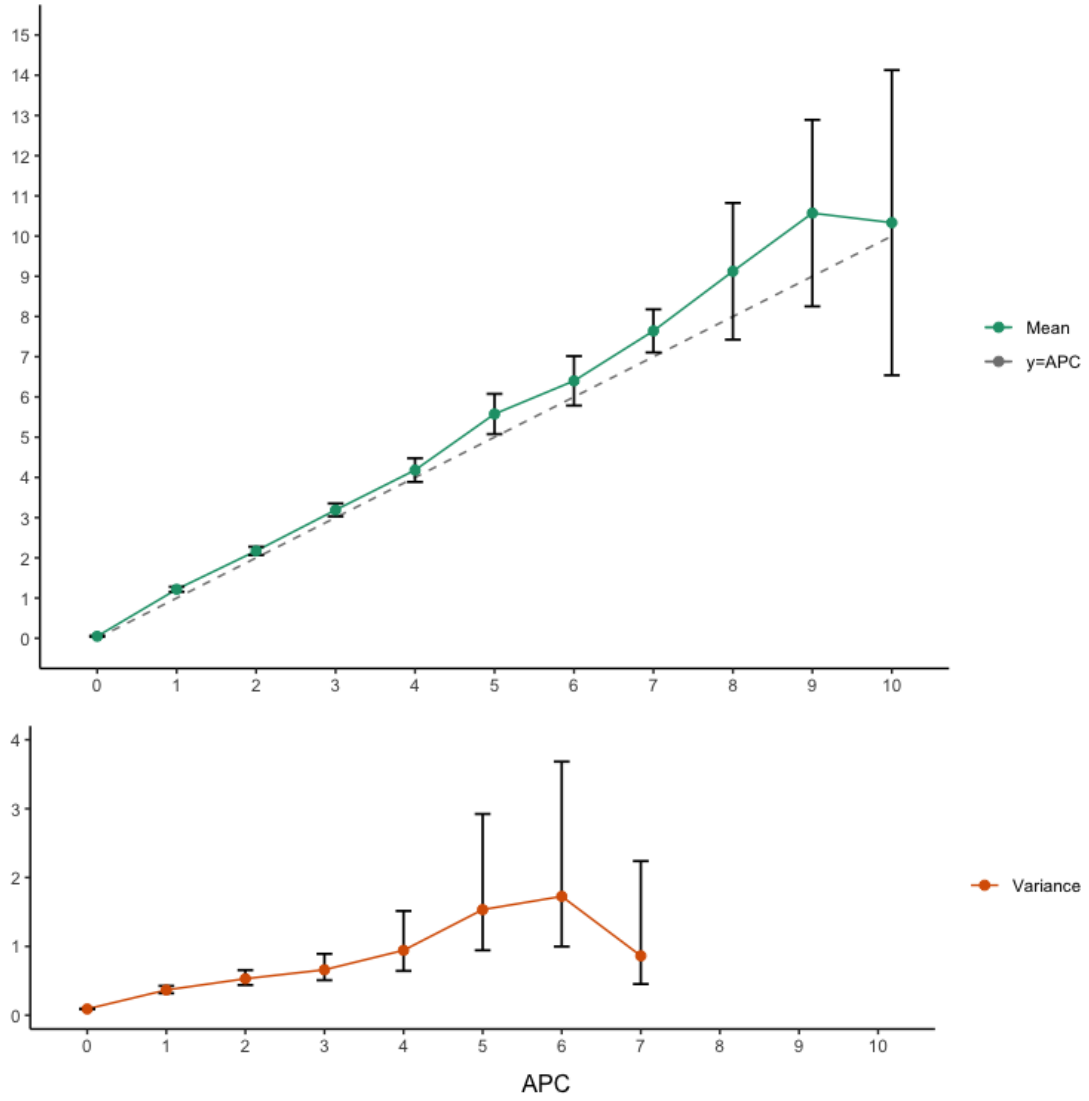


Figure 2.1: Empirical mean and variance of PC, with 95%-confidence intervals.

		Total PC			
Grouping Variable	**n**	**Min**	**Median**	**Mean**	**Max**
$stop$	686	0	1	3.02	48
$journey$	22	16	73	94.09	330

Table 2.5: Summary Statistics, stop- and journey-level.

insights from lower values to higher APC values as well. In Figure 2.1, we plot the mean and variance of PC as functions of APC, which reveals a linear relation for the mean. This suggests that a GLM for the expectation of PC could be a suitable option.

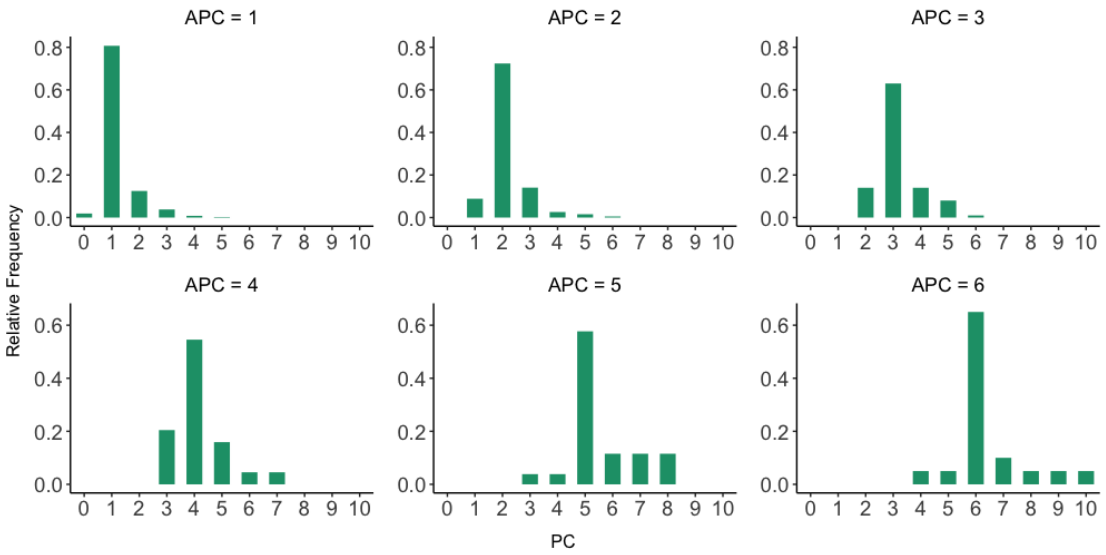
It is evident that the variance increases at a slower rate than the mean. When the variance is lower than the mean, the data exhibits underdispersion compared to the Poisson distribution. We might also notice a linear trend for the variance, suggesting a consistent level of underdispersion. Given the scarcity of data points in the upper range of APC, we only have reliable estimates of the variance for $APC \leq 7$.

In Figure 2.2a, we display the empirical, marginal distribution of PC, for APC values ranging from 1 to 6. There is a pronounced peak in $PC = APC$ which decreases in height as APC increases. The distribution around this peak is tight and seems slightly positively skewed. This indicates a tendency to undercount, confirmed by the fact that the mean is consistently above the dotted line $y = APC$ in Figure 2.1. This undercounting tendency is consistent with AtB's experiences.

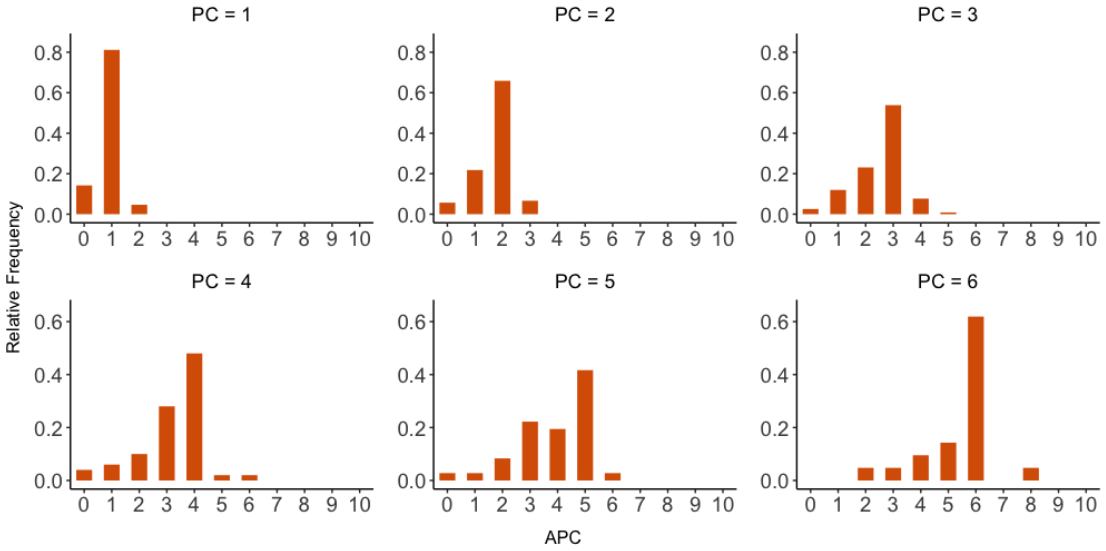
We then plot the empirical, marginal distribution of APC, given PC, in Figure 2.2b. In this figure, we observe a strong negative skew, clearly illustrating the tendency to undercount. The distribution resembles a triangle in the range $APC \leq PC$, which is reminiscent of a binomial distribution with parameter $n = PC$. A binomial distribution for APC would make sense as it would correspond to a straightforward DGP with a probability of counting each passenger. However, we notice that APC occasionally overcounts, so a binomial distribution alone would not be sufficient.

2.3.2 Stop- and Journey-Level

Stop- and journey-level data sets are created by grouping and aggregating door-level counts by the variables $stop$ and $journey$ respectively. Figure 2.3 displays the distribution of PC at stop- and journey-level, and how this differs between metro and ordinary buses. Table 2.5 presents some summary statistics for the total PC. From Figure 2.3a we note that there are around 300 stops where with no boarding passengers, most of which are from the routes 10, 11, 14, with bus_type ="ordinary". The bulk of the larger counts, $PC \in [5, 20]$, are from routes 1 and 3, with bus_type ="metro". Metro buses also accounts for almost all very large counts, $PC > 20$. These observations are further reflected in Figure 2.3b, which shows that journeys made by metro buses tend to have a higher number of total passengers. It is also noteworthy that there are two outliers that have a much higher total PC than both the median and the mean.

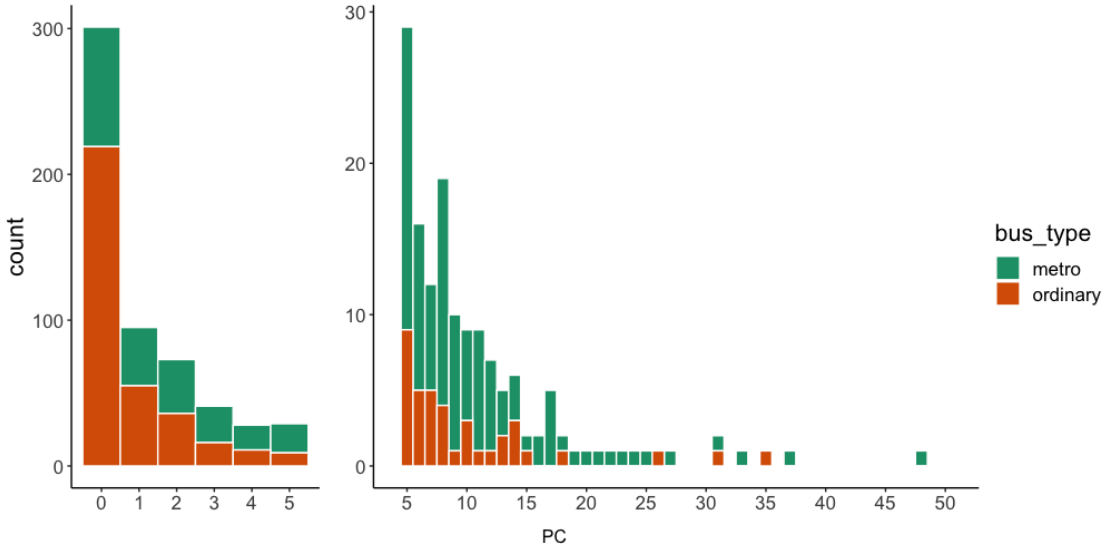


(a) Marginal distribution of PC, given APC.

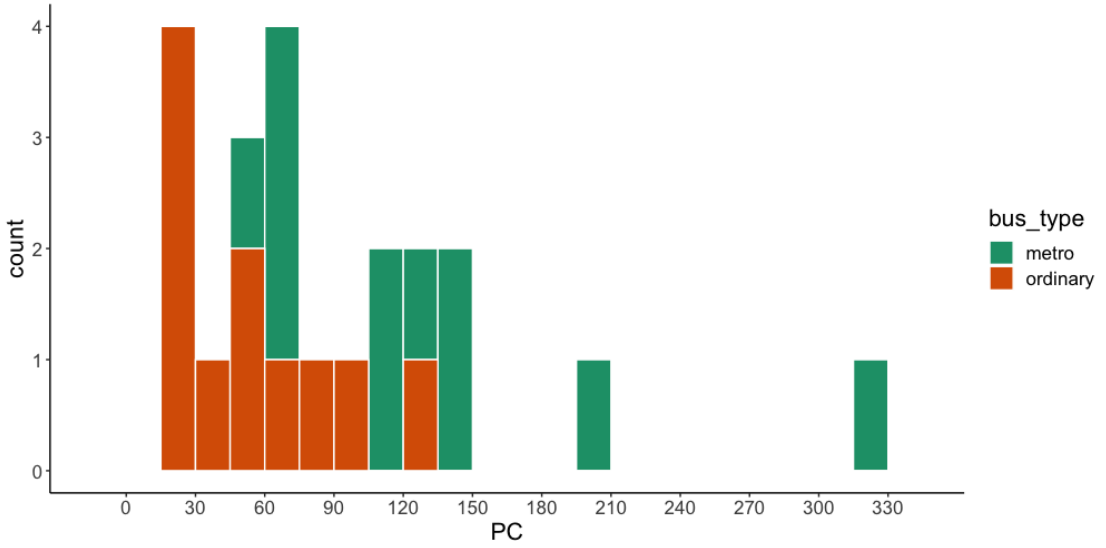


(b) Marginal distribution of APC, given PC.

Figure 2.2: Empirical distributions.



(a) Stops



(b) Journeys

Figure 2.3: Distribution of total PC at stop- and journey-level.

3

Background

This chapter presents the necessary theoretical background for the statistical methods and models which are employed in Chapter 4. The main bulk is attributed to count data, its characteristics, probability functions and linear count regression. We then give an account of maximum likelihood estimation and Bayesian inference. The last section concerns model assessment criteria, specifically the Akaike information criterion (AIC), (continuous) ranked probability score ((C)RPS) and probability integral transform (PIT) histograms.

3.1 Probability Functions for Count Data

Count data represents the number of occurrences of an event within a fixed interval. Illustrative examples are the number of patrons entering a store during a designated time frame, or as investigated in this thesis, the count of passengers boarding a bus at a particular stop. It is crucial to acknowledge the distinct characteristics of count data, as count random variables are inherently integer valued and non-negative. To effectively model and analyze count data, we utilize a range of count probability functions, tailored to its unique properties. Subsequent sections introduces a selection of these functions, which play a vital role in our analysis.

3.1.1 Poisson

If Y is Poisson distributed with parameter $\mu > 0$, then its probability mass function, PMF, is given by

$$P[Y = y \mid \mu] = \frac{(\mu)^y}{y!} e^{-\mu}, \qquad \text{for } y \in \{0, 1, 2, \ldots\}, \tag{3.1}$$

with expectation and variance

$$E[Y] = V[Y] = \mu.$$

The Poisson PMF involves a single parameter and is equidispersed, the mean and variance is always equal. (Cameron and Trivedi, 2013, p.3)

3.1.2 Binomial

A Bernoulli trial is a a trial with exactly two possible outcomes which we denote "success" and "fiasco". Let Y be a random variable representing the number of successes in a sequence of n independent Bernoulli trials, each with probability p of success. Y then follows the binomial distribution with PMF given by

$$\mathrm{P}[Y = y \mid n, p] = \left(\begin{array}{c} n \\ y \end{array} \right) p^y (1-p)^{n-y}, \qquad \text{for } y \in \{0, 1, \ldots, n\}, \tag{3.2}$$

where $0 < p < 1$. The expectation and variance of Y is given by

$$E[Y] = np, \quad V[Y] = np(1-p).$$

(Rigby et al., 2019, p.167,168)

3.1.3 Negative Binomial

The negative binomial distribution has several parametrizations, the most common of which is the NB2 (Cameron and Trivedi, 2013, p.81). The PMF of an NB2 distributed random variable, Y, is given by

$$P[Y = y \mid \mu, \sigma] = \frac{\Gamma\left(y + \frac{1}{\sigma}\right)}{\Gamma\left(\frac{1}{\sigma}\right)\Gamma(y+1)} \left(\frac{\sigma\mu}{1+\sigma\mu}\right)^y \left(\frac{1}{1+\sigma\mu}\right)^{1/\sigma}, \tag{3.3}$$

for $y \in \{0, 1, 2, \ldots\}$, $\mu > 0$ and $\sigma > 0$.

This is an equivalent parametrization to the one employed by Anscombe (1950), with the distinction being his use of $\alpha = 1/\sigma$ rather than σ (Rigby et al., 2019, p.483). The expectation and variance is given by

$$E[Y] = \mu, \quad V[Y] = \mu + \frac{1}{\sigma}\mu.$$

Thus $V[Y] \geq E[Y]$ and it follows that the NB distribution only allow for overdispersion.

3.1.4 Poisson-Inverse Gaussian

For long-tailed data the negative binomial distribution has limitations. The Sichel distribution is a distribution more suited for this type of data and a special case of the Sichel is the Poisson-inverse Gaussian (PIG) distribution. (Cameron and Trivedi, 2013, p.123,124)

The PMF of a PIG distributed random variable Y is given by

$$\mathrm{P}[Y = y \mid \mu, \sigma] = \left(\frac{2\alpha}{\pi}\right)^{1/2} \frac{\mu^y e^{1/\sigma} K_{y-\frac{1}{2}}(\alpha)}{y!(\alpha\sigma)^y}, \qquad \text{for } y \in \{0, 1, 2, \ldots\}, \tag{3.4}$$

where $\mu > 0$, $\sigma > 0$, $\alpha^2 = \sigma^{-2} + 2\mu\sigma^{-1}$ and $\alpha > 0$ (Rigby et al., 2019, p.487). $K_\lambda(t)$ is the modified Bessel function of the second kind given by

$$K_\lambda(t) = \frac{1}{2}\int_0^\infty x^{\lambda-1} \exp\left[-\frac{1}{2}t\left(x + x^{-1}\right)\right] dx$$

(Abramowitz (1965), as cited in Rigby et al. (2019, p.487)).

The expectation and variance of Y is given by

$$E[Y] = \mu, \quad V[Y] = \mu + \sigma\mu^2$$

(Rigby et al., 2019).

3.1.5 Double Poisson

The double Poisson distribution was introduced by Efron (1986) in his article on double exponential families. Double exponential families can be used to generalize any one parameter exponential family distribution to include an additional dispersion parameter in such a way that it enjoys exponential family properties for both parameters simultaneously.

The PMF of a double Poisson distributed random variable, Y, is given by

$$P[Y = y \mid \mu, \sigma] = \left(\frac{1}{\sigma}\right)^{1/2} e^{-\mu/\sigma} \left(\frac{e^{-y}y^y}{y!}\right) \left(\frac{e\mu}{y}\right)^{y/\sigma} \cdot C(\mu, \sigma) \tag{3.5}$$

for $y \in \{0, 1, 2, \ldots\}$, $\mu > 0$ and $\sigma > 0$. $C(\mu, \sigma)$ is a normalizing constant given by

$$C(\mu, \sigma) = \left[\sum_{y=0}^{\infty} \sigma^{-1/2} e^{-\mu/\sigma} \left(\frac{\mu}{y}\right)^{y/\sigma} \frac{e^{y/\sigma - y} y^y}{y!}\right]^{-1}.$$

The expectation and variance is approximately

$$E[Y] \approx \mu, \quad V[Y] \approx \mu \cdot \sigma, \tag{3.6}$$

these are very accurate approximations (Efron, 1986, p.715). As such the double Poisson distribution allows for both overdispersion, $\sigma > 1$, and underdispersion, $\sigma < 1$.

3.1.6 K-inflated Distributions

A common source of over- and underdispersion is the inflated presence of a certain value relative to the assumed distribution. Much consideration has been given to the class of models accounting for excess zeros. These models are known as zero-inflated models and are particularly useful in situations where zeros might arise from two different processes, one generating "true zeros" and the other generating non-zero values that can sometimes be zero. This is done by adding a binary component which inflates the probability of zero. For a count PMF, $f(y \mid \boldsymbol{\theta})$, the zero-inflated PMF is given by

$$\mathrm{P}[Y = y \mid \boldsymbol{\theta}, \nu] = \begin{cases} \nu + (1-\nu)f(y), & \text{if } y = 0 \\ (1-\nu)f(y) & \text{if } y > 0 \end{cases}, \quad \text{where } 0 < \nu < 1.$$

Recently some consideration has been given to the more general model where the PMF is inflated at one or several values $k \geq 0$ (Mohammadpour and Stasinopoulos

(2018), Arora et al. (2021), Arora and Chaganty (2021), Payandeh Najafabadi and Mohammadpour (2018)). In this thesis one such model, the k-inflated double Poisson, is given consideration. It inflates exactly one value, k, and has PMF

$$P[Y = y \mid \mu, \sigma, \nu] = \begin{cases} \nu + (1-\nu)f(k \mid \mu, \sigma), & \text{if } y = k \\ (1-\nu)f(y \mid \mu, \sigma) & \text{if } y \neq k \end{cases} \tag{3.7}$$

where $f(y \mid \mu, \sigma)$ is the PMF of the double Poisson distribution (3.5) and $0 \leq \nu \leq 1$.

3.2 Linear Count Regression

Considering a count response variable, Y, and a set of explanatory variables, $\boldsymbol{X} = [X_1, X_2, ..., X_p]$, combined in the linear predictor

$$\eta_i = \boldsymbol{x}'\boldsymbol{\beta} = \beta_0 + \beta_1 \cdot x_{i1} + \beta_2 \cdot x_{i2}, \ldots, \beta_p \cdot x_{ip}.$$

A linear count regression model assumes that Y follows a count distribution, such as Poisson or negative binomial. The mean, and potentially other parameters, are connected to the linear predictor through a link function, $g(\cdot)$ (Fahrmeir et al., 2013, p.293).

The classical linear model assumes normal distribution for Y, while linear count regression models fit within the broader framework of generalized linear models (GLMs) (Fahrmeir et al., 2013, p.269)

3.2.1 Link Functions

The selection of an appropriate link function is crucial, as it depends on the parameter domain, the covariate domain, and the assumed nature of their relationship. In the case of Poisson and other count distributions, the log-link function is frequently employed for the mean, μ,

$$log(\mu_i) = \eta_i. \tag{3.8}$$

Utilizing the log-link function ensures that μ remains non-negative and causes the covariates to have an exponential multiplicative effect on the expected value of Y. In cases where an additive effect is desired, one would use the direct relationship

$$\mu_i = \eta_i, \tag{3.9}$$

known as the identity link. To guarantee non-negativity when employing the identity link, it may be necessary to impose constraints on the parameter space of $\boldsymbol{X}$ and $\boldsymbol{\beta}$.

In some cases, it may be desirable to incorporate parameters such as the probability, p, in the Binomial and the weight, ν, in the k-inflated models. To accommodate this, a link function with a domain in the interval $[0, 1]$ should be selected for these parameters. The logit-link

$$\text{logit}\,(p_i) = \ln\left(\frac{p_i}{1-p_i}\right) = \eta_i \tag{3.10}$$

is a commonly used option.

3.3 Maximum Likelihood Estimation

For the set of n random variables $\boldsymbol{Y} = [Y_1, Y_2, \ldots, Y_n]$, with densities $f(y_i|\boldsymbol{\theta})$,

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} f(y_i|\boldsymbol{\theta})$$

is the joint likelihood of the realized data $\boldsymbol{y} = [y_1, y_2, \ldots, y_n]$. Now the maximum likelihood estimate (MLE), $\hat{\boldsymbol{\theta}}$, maximizes $L(\boldsymbol{\theta})$. In most situations working with the log-likelihood, $l(\boldsymbol{\theta}) = log(L(\boldsymbol{\theta}))$, is preferred as the joint log-likelihood can be expressed as a sum rather than a product. $l(\boldsymbol{\theta})$ and $L(\boldsymbol{\theta})$ share maximum due to log being a strictly monotonic function. (Givens and Hoeting, 2012, p.9)

3.4 Bayesian Inference

When doing Bayesian inference one view the parameters $\boldsymbol{\theta}$ of the likelihood $f(\boldsymbol{y}|\boldsymbol{\theta})$ as random variables. The prior distribution $f(\boldsymbol{\theta})$ is the density of $\boldsymbol{\theta}$ prior to observing the data. This prior knowledge is updated after observing the data $\boldsymbol{y} = [y_1, y_2, \ldots, y_n]$ using Bayes theorem:

$$f(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{f(\boldsymbol{\theta}) \cdot f(\boldsymbol{y}|\boldsymbol{\theta})}{f(\boldsymbol{y})}, \tag{3.11}$$

where $f(\boldsymbol{y}) = \int_\Theta f(\boldsymbol{\theta}) \cdot f(\boldsymbol{y}|\boldsymbol{\theta}) d\boldsymbol{\theta}$. $f(\boldsymbol{\theta}|\boldsymbol{y})$ is the posterior distribution and represents the density of $\boldsymbol{\theta}$ after observing the data. (Givens and Hoeting, 2012, p.11)

For a single, discrete parameter θ, (3.11) can be rewritten as

$$f(\theta|\boldsymbol{y}) = \frac{f(\theta) \cdot f(\boldsymbol{y}|\theta)}{\sum_\Theta f(\theta) \cdot f(\boldsymbol{y}|\theta)}. \tag{3.12}$$

3.4.1 Empirical Bayes

In traditional Bayesian statistics, the prior distribution, $f(\theta)$, is chosen based on prior knowledge. However, in many practical situations, this prior knowledge may not be readily available or might be hard to quantify. The empirical Bayesian estimates the prior from the data itself, and can often obtain better results (Casella, 1985).

3.5 Model Assessment

In order to differentiate between models, certain criteria must be established. The Akaike information criterion (AIC) offers a quantitative criteria by which the model goodness-of-fit can be evaluated. The Continuous Ranked Probability Score (CRPS) allows us to gauge the accuracy of probabilistic predictions and the Probability Integral Transform (PIT) histograms provides a visual tool to assess the calibration.

3.5.1 AIC

The Akaike information criterion (AIC) can be used as a criteria to compare non-nested models. AIC is among the most popular model choice criteria (Fahrmeir et al., 2013, p.148) and is defined as

$$AIC = -2log(L(\boldsymbol{y})) + 2k, \tag{3.13}$$

where $L(\boldsymbol{y})$ is the likelihood of the observed data $\boldsymbol{y}$ for the given model, and k is the number of estimated parameters. A lower AIC value is preferred as this suggests good model fit through a larger log-likelihood while penalizing model complexity through k.

The AIC is easily calculated when doing maximum likelihood estimation as the likelihood is already available.

3.5.2 (C)RPS

The Ranked Probability Score (RPS) is a scoring rule used to evaluate the accuracy of probabilistic predictions. It is used for ordinal predictions, where the possible outcomes have a natural successive order. It is a strictly proper scoring rule, meaning the score cannot be improved by trying to hedge the prediction (Wilks, 2011, p.418). Among strictly proper scoring rules, RPS is by far the most popular (Wilks, 2011, p.420). Continuous ranked probability score (CRPS) is the extension of RPS to continuous outcomes.

The intuition for (C)RPS is that it is a measure of the deviation from a perfect prediction. A perfect prediction would predict the observed outcome with absolute certainty, probability equal 1. If we express our predictions as cumulative density functions (CDFs), the CDF of the perfect prediction is the unit step function at the observation, we call this the cumulative observation. Now RPS is the squared difference between the prediction model CDF and the cumulative observation. A lower score is better for both RPS and CRPS, rewarding a probabilistic prediction that is concentrated about the observed value. This is illustrated for continuous outcomes in Figure 3.1, where the better prediction model has a CDF closer to the cumulative observation, resulting in a lower CRPS.

When evaluating a set of n observations we simply average the (C)RPS values,

$$\overline{\text{(C)RPS}} = \frac{1}{n}\sum_{k=1}^{n} \text{(C)RPS}_k. \tag{3.14}$$

RPS

Following the notation of Wilks (2011) let $[1, \ldots, M]$ be the set of possible outcomes with prediction probabilities $\boldsymbol{y} = [y_1, \ldots, y_M]$. The cumulative prediction is defined to be the vector

$$\boldsymbol{Y} = [Y_1, \ldots, Y_M],$$

where

$$Y_j = \sum_{m=1}^{j} y_m, \; \mathrm{j} \in [1, \ldots, M].$$

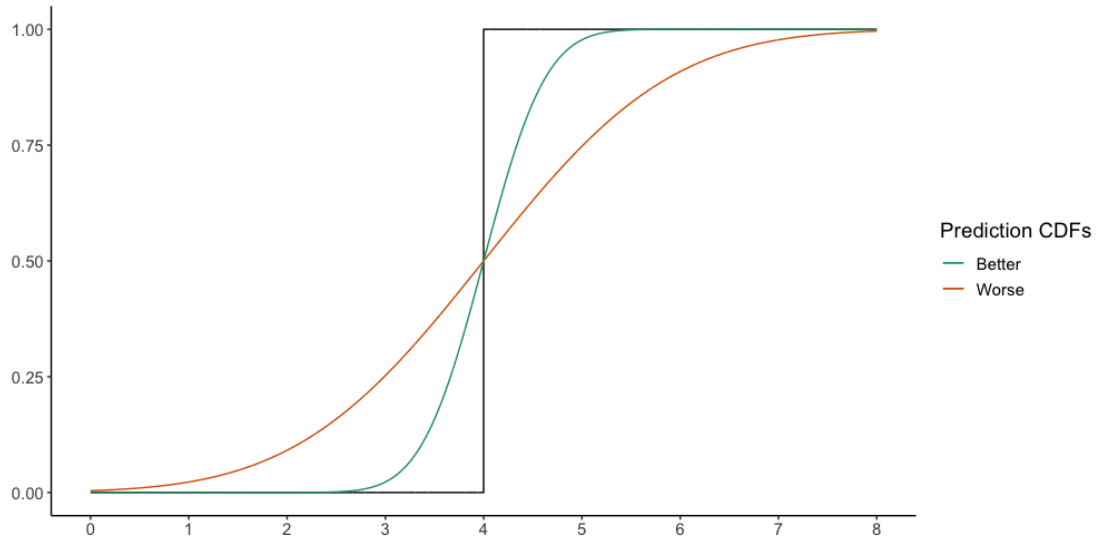


Figure 3.1: CRPS illustration.

The observation vector is defined as $\boldsymbol{o} = [o_1, \ldots, o_M]$, where for observed outcome k, $o_j = 1$ for $j = k$ and $o_j = 0$ for $j \neq k$. We define the cumulative observation to be the vector

$$\boldsymbol{O} = [O_1, \ldots, O_M],$$

where

$$O_j = \sum_{m=1}^{j} o_m, \ \mathrm{j} \in [1, \ldots, M].$$

The RPS is defined as

$$RPS = \sum_{j=1}^{M} (Y_j - O_j)^2. \tag{3.15}$$

CRPS

Continuous ranked probability score (CRPS) is the extension of RPS to continuous outcomes. Again following the notation of Wilks (2011), the prediction is now given as a a probability density function (PDF) $f(y)$ with cumulative density function (CDF) $F(y)$. Define the cumulative observation as the step function

$$F_o(y) = \begin{cases} 0, & \text{for } y < o \\ 1, & \text{for } y \geq o, \end{cases}$$

when the observed outcome is $y = o$. CRPS is then defined as

$$CRPS = \int_{-\infty}^{\infty} [F(y) - F_o(y)]^2 dy. \tag{3.16}$$

3.5.3 PIT

The probability integral transformation (PIT) theorem states that for a random variable Y with continuous CDF $F_Y()$, $U = F_Y(Y)$ is uniformly distributed on $(0, 1)$ (Casella

and Berger, 2002). This property can be used to create PIT histograms, which allows us to visually evaluate the calibration of a modeled CDF $F'()$. The procedure is as follows:

- For observed data $\boldsymbol{y}$, transform to PIT residuals

$$\boldsymbol{u} = F'(\boldsymbol{y}). \tag{3.17}$$

- Plot a histogram of $\boldsymbol{u}$.
- The PIT histogram should resemble that of the uniform distribution on $(0, 1)$ if $F'() = F_Y()$.

When $F_Y()$ is discrete, the PIT theorem does not hold and we need a more general definition of the PIT residuals. Following Dunn and Smyth (1996) we let

$$a_i = \lim_{y \to y_i} F_Y(y) \quad \text{and} \tag{3.18}$$

$$b_i = F_Y(y_i), \tag{3.19}$$

and define the randomized quantile residuals as

$$r_i = \Phi^{-1}(u_i), \tag{3.20}$$

where u_i is a random variable that is uniformly distributed on $(a_i, b_i]$. Now r_i are exactly standard normal (Dunn and Smyth, 1996) irrespective of the exact values of a_i and b_i. It follows from the PIT theorem that u_i are uniformly distributed on (0,1).

A limitation of PIT histograms is that they don't differentiate between observations with extreme PIT residuals, close to 0 or 1, and outright outliers (Dunn and Smyth, 1996).

4

Probabilistic Models and Methods for PC

In this chapter we propose two different door-level, probabilistic models for PC conditional on APC. The first is a GLM, with APC as explanatory variable, for which we consider the Poisson, double Poisson and k-inflated double Poisson distributions, the latter two to account for the underdispersion. For the second model, ***DGP-M***, we seek to model the true data generating process (DGP) and we make inference on PC using empirical Bayes. The chapter starts by presenting the models, and the choices made during the modeling procedure. We then discuss how the door-level models are applied to stop- and journey-level aggregates. The chapter concludes with a section on model evaluation and an account of the software used.

4.1 Generalized Linear Model - *GLM*

The EDA in Section 2.3 revealed a linear relation between APC and the mean of PC. Ignoring the fact that the true DGP clearly operates in the opposite direction, a logical approach is a GLM for PC, with APC as explanatory variable, and a count distribution for PC. By fitting separate parameter estimates for the scenario in which $APC = 0$, and by assuming $a_0 = 0$ and $a_1 > 1$, we can ensure a strictly positive predictor,

$$\eta_{\mu,i} = a_0 + a_1 \cdot apc_i. \tag{4.1}$$

A positive predictor allows the use of identity link which lets us model an additive effect of APC on the expected value of PC. We use the predictor (4.1) with identity link (3.9) for the mean parameter of all considered distributions.

4.1.1 Poisson

PC is a count variable, limiting our choice of probability distribution. The "simplest and most widely used choice" (Fahrmeir et al. (2013), p. 293) for count variables is the Poisson distribution. The Poisson-PMF (3.1) is equidispersed, the mean and variance is equal. This equidispersion property of the Poisson is often too restrictive, and this is in fact the case for our data set. As found during the EDA the conditional variance is smaller than the conditional mean by a considerable amount (2.1). PC, given APC, is therefore underdispersed relative to the Poisson-PMF.

When the Poisson proves to be too restrictive, it is natural to consider more flexible models such as the negative binomial which breaks the equidispersion restriction by introducing an extra parameter (Cameron and Trivedi, 2013, p.18). However most of the popular alternative models, including the negative binomial, only account for overdispersion and offers no improvement on the Poisson for underdispersed data. In this thesis the Poisson distribution is used as a benchmark. The double Poisson and k-inflated double Poisson distributions are considered to account for the underdispersion relative to the Poisson.

4.1.2 Double Poisson

Most research seem to tackle the issue of overdispersion and this is understandable since, as Cameron and Trivedi (2013, p.169) note, "overdispersion is far more common than underdispersion." Still there are some two-parameter extensions of the Poisson distribution that accommodate underdispersed counts. The generalized Poisson distribution theoretically allows for underdispersion (Hilbe, 2014, p.211); however, all R implementations encountered were limited to handling overdispersion only. The two other options of note are the Conway-Maxwell-Poisson (CMP) distribution and the double Poisson distribution, neither of which has been investigated much until recently (Zou et al., 2013, p.498). Zou et al. (2013) found CMP and double Poisson to be comparable in terms of goodness-of-fit. double Poisson is our choice as it is implemented in the R package $gamlss.dist$ (Stasinopoulos and Rigby, 2022).

The double Poisson-PMF (3.5) has an additional dispersion parameter which can be modeled separately from the mean. This should allow us to obtain a better fit. The variance is approximately $\sigma \cdot \mu$ (3.6), so $\sigma < 1$ gives underdispersion relative to the Poisson. We use the log-link (3.8) for the dispersion parameter σ. Both a model with constant level of underdispersion,

$$\eta_{\sigma,i} = b_0, \tag{4.2}$$

and a model with linear predictor,

$$\eta_{\sigma,i} = b_0 + b_1 \cdot apc_i, \tag{4.3}$$

were considered. The inclusion of b_1 was found to be insignificant, so we exclude it going forward.

4.1.3 K-inflated Double Poisson

We found in the EDA that in the empiric marginal distributions of PC, given APC, there is a peak in frequency of $PC = APC$. To include this property of the data we consider the k-inflated double Poisson distribution (3.7). This distribution is useful when there is inflation of a specific value, k, relative to the double Poisson.

Research has been conducted on k-inflated count models within regression frameworks (Arora and Chaganty (2021), Arora et al. (2021), Payandeh Najafabadi and Mohammadpour (2018)). In all the studies reviewed, the inflation point, k, remained constant for all predictor values. However, our unique situation calls for the incorporation of k as a parameter in the model.

To achieve this we propose a model where the inflation point is directly determined by the APC value:

$$k_i = apc_i.$$

The weight put on the inflation is determined by the parameter ν, and it is modeled through a logit-link (3.10) with the linear predictor

$$\eta_{\nu,i} = c_0 + c_1 \cdot apc_i.$$

4.1.4 Zero-Counts

As noted in Section 2.3, $APC = 0$ makes up about 71% of the data set and is correct about 96% of the time. We therefore have concerns that the model would lean too heavily on fitting the zeros correctly, if they were included. The choice was made to separate out $APC = 0$ and estimate the parameters separately. As we have already mentioned, this also allows for the use of identity-link (3.9) for μ.

For the zero-counts the choice of link functions are less complicated as we don't have any covariates to worry about. We use the log-link (3.8) for μ and σ, ensuring non-negativity, and logit-link (3.10) for ν, as $0 < \nu < 1$.

4.1.5 Summary of Mathematical Model Specifications

Three GLMs are considered for $PC_i \mid apc_i$. We utilize the identity-link with intercept $a_0 = 0$ for the mean parameter, μ, for all models. A single estimate is used for the dispersion parameter, σ, of the double Poisson and k-inflated double Poisson. A linear logit-link is used for the inflation weight parameter, ν, of the k-inflated double Poisson. When $apc_i = 0$, the log-link (3.8) is used for μ and σ, and the logit-link (3.10) for ν.

Poisson - *PO*

$$PC_i \mid apc_i \sim Poisson(\mu_i) \tag{4.4}$$
$$\mu_i = \begin{cases} exp(a_{zero}), & \text{if } apc_i = 0 \\ a_1 \cdot apc_i, & \text{if } apc_i \neq 0 \end{cases}$$

Double Poisson - *DPO*

$$PC_i \mid apc_i \sim DPO(\mu_i, \sigma_i) \tag{4.5}$$
$$\mu_i = \begin{cases} exp(a_{zero}), & \text{if } apc_i = 0 \\ a_1 \cdot apc_i, & \text{if } apc_i \neq 0 \end{cases}$$
$$\sigma_i = \begin{cases} exp(b_{zero}), & \text{if } apc_i = 0 \\ exp(b_0), & \text{if } apc_i \neq 0 \end{cases}$$

K-inflated Double Poisson - *KIDPO*

$$
\begin{aligned}
PC_i \mid apc_i &\sim KIDPO(\mu_i, \sigma_i, \nu_i, k_i) \\
\mu_i &= \begin{cases} exp(a_{zero}), & \text{if } apc_i = 0 \\ a_1 \cdot apc_i, & \text{if } apc_i \neq 0 \end{cases} \\
\sigma_i &= \begin{cases} exp(b_{zero}), & \text{if } apc_i = 0 \\ exp(b_0), & \text{if } apc_i \neq 0 \end{cases} \\
\nu_i &= \begin{cases} \frac{exp(c_{zero})}{1+exp(c_{zero})}, & \text{if } apc_i = 0 \\ \frac{exp(c_0+c_1 \cdot apc_i)}{1+exp(c_0+c_1 \cdot apc_i)}, & \text{if } apc_i \neq 0 \end{cases} \\
k_i &= apc_i
\end{aligned}
\tag{4.6}
$$

4.2 Data Generating Process Model - *DGP-M*

We also propose an alternative to the GLMs considered thus far, which focuses on modeling the APC counting process. By doing so, we can better represent the true DGP and obtain a model for APC conditional on PC. Still our primary objective remains the development of a model for PC conditional on APC.

To accomplish this, we assume a distribution for PC and apply Bayes' theorem (3.11), effectively treating the distribution of PC as a prior distribution in a Bayesian context. We have no strong beliefs or knowledge about the distribution of PC, and the modeling of passenger flow is beyond the scope of this thesis. We therefore choose to use an empirical prior on PC, based on the observed data. It is crucial to emphasize that our approach is predominantly frequentist in nature, as opposed to a pure Bayesian methodology that would place priors on all included parameters, thereby incorporating the uncertainty from these parameters throughout the model.

The proposed model takes the form of Bayes theorem:

$$
f(pc|apc) = \frac{f(pc) \cdot f(apc|pc)}{\sum_{PC} f(pc) \cdot f(apc|pc)} \tag{4.7}
$$

Where the likelihood function, $f(apc \mid pc)$, is a model for the DGP, and $f(pc)$ is the prior on PC. Both the prior and the likelihood is fit individually by maximum likelihood. Then if we assume a finite number of possible values for PC, we can calculate the normalizing constant fairly easily.

4.2.1 Likelihood function

The first step is a model for the likelihood function $f(apc_i \mid pc_i)$. If X is a measurement with error of the true varaible Y, the typical way to model X is too assume $X = Y + \epsilon$, where Y is a deterministic component and ϵ is a (normal) random variable. This is fine for continuous variables, but count variables are non-negative. Therefore ϵ would have to be non-negative and the model can only describe overestimation of the true variable. In our case APC can be both smaller and larger than PC, so we need an alternative

approach. Cameron and Trivedi (2013, p.486) propose a model for the measurement error of a random count variable which allows for both undercounting and overcounting. Using our variable names: APC is assumed a sum of a binomial distributed variable with a probability, p, of counting a passenger and number of boarding passengers, $n = PC$; and a Poisson distributed variable with mean μ. The binomial component models the undercount and the Poisson the overcount.

$$\begin{aligned} APC_i \mid pc_i &= X + Y, \\ \text{where } X &\sim Bi(p_i, n_i = pc_i) \\ \text{and } Y &\sim Po(\mu_i). \end{aligned} \tag{4.8}$$

The calculation of $f(apc_i \mid pc_i)$ can be done easily as the convolution of the two distributions. Let $f_X(x \mid pc_i)$ denote the PMF of X and $f_Y(y)$ the PMF of Y. Then

$$\begin{aligned} f(apc_i \mid pc_i) &= \sum_{m=0}^{apc_i} \boldsymbol{X}[m] \cdot \boldsymbol{Y}[apc_i - m], \\ \text{where } \boldsymbol{X} &= [f_X(0 \mid pc_i) \quad , f_X(1 \mid pc_i), \ldots], \\ \boldsymbol{Y} &= [f_Y(0) \qquad , f_Y(1) \qquad , \ldots]. \end{aligned} \tag{4.9}$$

The mean parameter of Y, μ, is modeled with the log-link (3.8) and is assumed to be constant for all $pc_i \geq 1$. We use the logit-link (3.10) for the parameter p of X, and also assume it to be constant.

When there are no boarding passengers, $PC = 0$, only overcounting is possible. It could therefore be natural to assume the same overcounting error for both $pc_i = 0$ and $pc_i \geq 1$. However these are quite different situations. When $pc_i = 0$ the door might not even have opened at all. We also noted in Section 2.3 that when there are no boarding passengers, overcounting only occurred 7 out of 1904 times. We choose to fit separate estimates for μ for $PC = 0$ and $PC \neq 0$.

The parameters of (4.8) is modeled as follows:

$$\begin{aligned} p_i &= \frac{exp(p_0)}{1 + exp(p_0)} \\ \mu_i &= \begin{cases} exp(b_{zero}), & \text{if } pc_i = 0 \\ exp(b_0), & \text{if } pc_i \neq 0 \end{cases} \end{aligned} \tag{4.10}$$

4.2.2 Prior

In order to obtain a model for PC conditional on APC, we also need a prior distribution on PC. We take an empirical Bayesian approach in which we estimate the parameters of this distribution from the data. Because there are no covariates, it is fairly straight forward to fit a range of count distributions by maximum likelihood, and select the best fit by AIC.

The zero-inflated Poisson-inverse Gaussian (ZIPIG) distribution is found to be the best fit, with parameters

$$\mu = 1.5081, \quad \sigma = 1.1811 \quad \text{and} \quad \nu = 0.50122.$$

The empirical prior has PMF:

$$f(pc \mid \mu, \sigma, \nu) = \begin{cases} \nu + (1-\nu)f(y \mid \mu, \sigma), & \text{if } y = 0 \\ (1-\nu)f(y \mid \mu, \sigma) & \text{if } y \neq 0 \end{cases} \tag{4.11}$$

where $f(y \mid \mu, \sigma)$ is the PMF of the Poisson-inverse Gaussian distribution (3.4) and $0 \leq \nu \leq 1$.

4.2.3 Normalizing Constant

Having specified and found parameter estimates for the prior and likelihood, the final piece is the normalizing constant in the denominator of (4.7). While this sum in theory goes to infinity, in practice we can safely set an upper limit on the possible values of PC. It is clear that there is a limit to how many passenger boards through one door during a stop. We noted in Section 2.3, that the maximum value of PC in our data set is 16, and that the frequency of observations decrease as PC increases. We conservatively set the upper limit to 25. Now the normalizing constant is a finite sum, and the calculations needed to find the values of $f(pc|apc)$ are simple and executed quickly in R.

4.3 Aggregate Models

The following subsections describe how we apply the proposed door-level models to stop- and journey-level aggregates.

4.3.1 Stop Level

There are four doors on a bus, so during a stop we have four door-level counts. We denote the random variable for the total number of boarding passengers at a single stop PC_{stop}. It is the sum of four, assumed independently distributed, random variables.

$$PC_{stop,i} = PC_{1,i} + PC_{2,i} + PC_{3,i} + PC_{4,i}.$$

At a stop we have the set of observations

$$\boldsymbol{apc}_i = [apc_{1,i}, apc_{2,i}, apc_{3,i}, apc_{4,i}].$$

Given the set of observations we from our models calculate the probability vectors

$$\boldsymbol{PC}_{j,i} = [P(PC_{j,i} = 0), P(PC_{j,i} = 1), \ldots, P(PC_{j,i} = 25)]$$

for each door $j \in [1, 2, 3, 4]$, assuming PC only takes values in $[0, 25]$. This assumption has already been made for *DGP-M*. The probability vectors from *GLM* nearly always sum to one, but are still normalized. Now the probability vector for $PC_{stop,i}$ can be calculated by repeated convolution.

$$\begin{aligned} \boldsymbol{PC}_{stop,i} &= [[[\boldsymbol{PC}_{1,i} * \boldsymbol{PC}_{2,i}] * \boldsymbol{PC}_{3,i}] * \boldsymbol{PC}_{4,i}] \\ &= [P(PC_{stop,i} = 0), P(PC_{stop,i} = 1), \ldots, P(PC_{stop,i} = 100)]. \end{aligned}$$

The models *GLM* and *DGP-M* are both applied in such a way to clusters of four door-level observations, grouped by the variable $stop$. The resulting probability vectors are considered discrete probabilistic models with possible outcomes $[0, 1, \ldots, 100]$.

4.3.2 Journey Level

There are 22 distinct journeys in our data set, these have between 96 and 164 observations at door-level 2.5. Assuming that these observations are independently distributed, the total number of boarding passengers for a single journey, denoted as $PC_{journey}$, follows a normal distribution, by the central limit theorem. The expected value and variance of $PC_{journey}$ can be calculated as the sum of the individual expectations and variances, respectively, for each door-level observation.

The individual expectations and variances at door-level,

$$E[PC_i] = \sum_{PC} pc \cdot f(PC_i = pc)$$

and

$$\begin{aligned} V[PC_i] &= E[PC_i^2] - (E[PC_i])^2 \\ &= \sum_{PC} [pc^2 \cdot f(PC_i = pc)] - (\sum_{PC} pc \cdot f(PC_i = pc))^2, \end{aligned}$$

are calculated numerically for both *GLM* and *DGP-M*. Theoretical expressions for the approximate expectation and variance of the k-inflated double Poisson-distribution were derived, but these deviated too much from the numerical calculations, especially for smaller values. The derived theoretical approximations can be found in Appendix A.

Door-level observations are grouped by the variable $journey$ and the total expectations and variances are calculated for both *GLM* and *DGP-M*. This results in continuous normal probabilistic models for $PC_{journey}$.

In addition to grouping door-level observations by $journey$ we create random groupings of 100 observations, we denote these groupings "random journeys". Four sets of random journeys are generated and provide contrast to the actual journeys. This allows us to investigate whether there are any correlation between door-level observations in the true journeys which are not present for the random journeys.

4.4 Model Evaluation

We now have four probabilistic models for the number of boarding passengers at door-level, PC, given the corresponding automatic count, APC. These are *PO*, *DPO*, *KIDPO* and *DGP-M*. We first compare *PO*, *DPO* and *KIDPO* and denote the preferred one *GLM*. *GLM* is then compared to *DGP-M*. We also investigate the performance of *GLM* and *DGP-M* on PC_{stop} and $PC_{journey}$.

In assessing the various models, we employ several statistical measures. At the door-level, the models are compared using AIC, RPS and PIT histograms. Each of these provide unique insights. AIC (3.13) measures the goodness-of-fit of the models to the training data, RPS (3.15) evaluates the accuracy of the probabilistic predictions, and PIT histograms visually presents the distribution of the PIT residuals (3.17), which is ideally close to the uniform distribution on $[0, 1]$, thus allowing us to assess the calibration of the models.

Further, at both the stop- and journey-level, the *GLM* and the *DGP-M* are compared using CRPS (3.16) and PIT histograms. The CRPS generalizes the RPS to the case of

continuous outcomes. PIT histograms again provide a visualization for assessing the calibration.

4.5 Software and Implementation

All data analysis and implementation is performed in R (R Core Team, 2021), through extensive use of the $tidyverse$ (Wickham et al., 2019) and $gamlss$ (Rigby and Stasinopoulos, 2005) packages. Plots are created using $ggplot2$ (Wickham, 2016).

The *PO*- and *DPO*-models are fitted using the $gamlss()$-function from the $gamlss$ package for R, with $method = RS()$ (Rigby and Stasinopoulos, 2005). It finds MLEs for the parameters using the RS algorithm which is thoroughly explained in Stasinopoulos et al. (2017, p.59-69).

Our proposed *KIDPO*-model expands on the existing framework by allowing k to vary as a function of the explanatory variable. No current implementations were found to allow for this novel approach. The $gamlss$-framework for example is only able to fit k-inflated models with fixed k. We therefore implement the log-likelihood for model (4.6) ourselves in R using $dKIDPO()$ from $gamlss.countKinf$ (Mohammadpour and Stasinopoulos, 2018). To find MLEs for the parameters of the *KIDPO*-model, we employ the $optim()$-function from base R. The log-likelihood is maximized directly using box-constrained quasi-newton, L-BFGS-B.

MLEs for the parameters in the likelihood function in *DGP-M*, are found similarly. (4.9) is implemented in R using $dpois()$ and $dbinom()$ which are included in base R. This is again maximized using $optim()$.

To find the best fit for the empirical prior, we make use of the $fitDist(type = "counts")$-function from $gamlss$ (Rigby and Stasinopoulos, 2005). This function fits all relevant parametric, count distributions, by maximum likelihood and presents their AIC.

AIC (3.13) is provided for all fits made using $gamlss()$, and are otherwise easily calculated from the log-likelihood. RPS (3.15) is calculated using $rps()$ from the $verification$-package (Laboratory, 2015), and for CRPS (3.16) we make use of $crps.numeric()$ from the package $scoringRules$ (Jordan et al., 2019).

4.5.1 PIT

The PIT residuals (3.17) represents the likelihood that a random variable is less than or equal to the observed value, according to the model. These are easily calculated for journey-level aggregates as the distributions are continuous.

For door- and stop-level however, our predictive distributions are discrete, and we must adjust for this so that the PIT histogram appears uniform for a well-fitting model. As presented in Subsection 3.5.3, the PIT residuals u_i are random variables for discrete distributions, uniformly distributed on $(a_i, b_i]$ (3.18) (3.19). Instead of computing a single PIT residual for each observation, we repeatedly sample u_i. This process generates a collection of PIT values for each observation. Subsequently, we merge these sets into a single vector, from which we construct the histogram. This approach enables us to account for the discreteness of our predictive distributions, ensuring that a uniform PIT histogram corresponds to a reliable model.

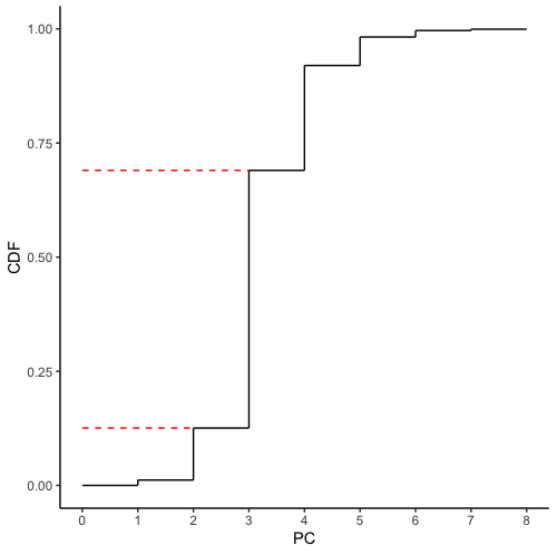


Figure 4.1: Discrete PIT illustration.

The approach can be illustrated using Figure 4.1 which shows a discrete, predictive CDF. The upper red line shows $b_i = 0.690$ and the lower red line shows $a_i = 0.126$. For an observation $pc_i = 3$ we repeatedly sample from the uniform distribution on the interval $(0.126, 0.690]$, between the red lines. 100 PIT samples are generated for each observation.

5

Results

In this chapter we present and compare the results for the probabilistic models for PC described in Chapter 4. We show parameter estimates and fitted PMFs, and present AIC- and (C)RPS-values as well as PIT histograms. We begin by considering the results of four door-level models for PC, first the three generalized linear models (GLMs), and then the data generating process model *DGP-M*. We then move on to the results for stop- and journey-level aggregates, PC_{stop} and $PC_{journey}$.

5.1 Results for Door-Level Probabilistic Models for PC

This section considers and compares the three GLMs; *PO*, *DPO* and *KIDPO*; both in terms of AIC and RPS, and visually by inspection of the respective PMFs and PIT histograms.

5.1.1 Generalized Linear Model - *GLM*

The three GLMs are presented in Table 5.1 with estimated predictors and AIC values, and parameter estimates and AIC values for zero-counts, $APC = 0$, in Table 5.2. The total AIC is presented along with RPS in Table 5.3. The fitted PMFs are plotted for some selected values of APC in Figure 5.1, and together over the empirical distributions in Figure 5.2. The PIT histograms are shown in Figure 5.3.

As expected, due to the *PO* being equidispersed, the resulting model has too much variance. When plotted over the empirical distribution in Figure 5.2 it is obvious that it is too flat. This is also confirmed by the PIT histogram (5.3) which shows a clear peak in the middle.

The *DPO* was included to improve on this shortcoming of the *PO*, and it is able to do so as we can see from the AIC value and in Figure 5.1. It has a much tighter distribution about the mean, and in Figure 5.2 it is clear that it is a better fit to the data. The PIT histogram of the *DPO* is an obvious improvement but the asymmetry is an indication that there is some skew in the data relative to the model predictions. There is also still an excess of observations at $PC = APC$ relative to the *DPO*, especially for larger values of APC.

The *KIDPO* was implemented in an effort to capture this peak in the empirical data. In Figure 5.1 it is clear that the model behaves like we wanted. The PMF has a peak

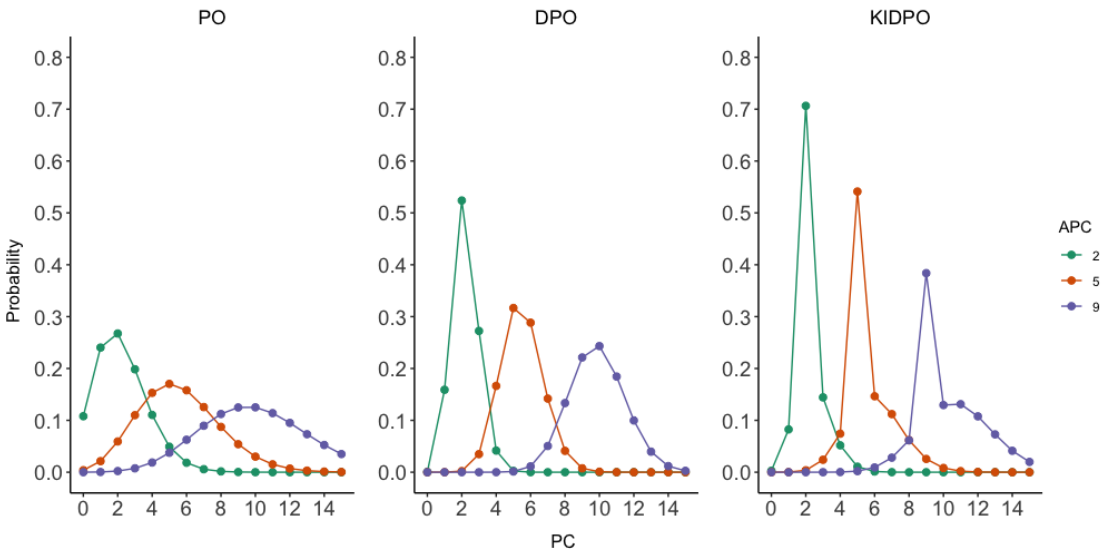


Figure 5.1: PMFs of the three GLMs, each shown for three selected values of APC.

in $PC = APC$ the weight of which decrease linearly with APC. The k-inflation also allows the double Poisson component to be shifted, as seen by the steeper $\eta_{\mu,i}$ (5.1). This lets the model reflect the skewed nature of the data. Figure 5.2 seem to indicate a much better fit to the data. This is confirmed by a considerably smaller AIC value (5.1) and the PIT histogram is now very close to uniform.

For zero-counts the situation is different as we don't have underdispersion. In fact we see that *DPO* and *KIDPO* fit a large overdispersion, $\sigma > 1$. Nevertheless, *KIDPO* is the best fit also here with a smaller AIC than the two others.

KIDPO is the best fit among the considered distributions in terms of AIC and PIT histograms, for $APC \neq 0$ and also for $APC = 0$. RPS values (5.3) backs this conclusion, with *KIDPO* having the smallest average RPS among the GLMs. We select it as our preferred GLM, and denote it *GLM*.

Model	$\boldsymbol{\eta_{\mu,i} = \mu_i}$	$\boldsymbol{\eta_{\sigma,i} = log(\sigma_i)}$	$\boldsymbol{\eta_{\nu,i} = logit(\nu_i)}$	**AIC**
PO	$1.113 \cdot apc_i$			2225.7
DPO	$1.097 \cdot apc_i$	-1.323		1625.4
KIDPO	$1.196 \cdot apc_i$	-0.854	$0.350 - 0.143 \cdot apc_i$	1496.3

Table 5.1: The three GLMs with estimated predictors; $\eta_{\mu,i}$, $\eta_{\sigma,i}$ and $\eta_{\nu,i}$; and AIC values for $apc_i \geq 1$.

Model	$\boldsymbol{\mu_i}$	$\boldsymbol{\sigma_i}$	$\boldsymbol{\nu_i}$	**AIC**
PO	0.049836			835.536
DPO	$2.036762 \cdot 10^{-16}$	16.57673		764.181
KIDPO	0.003997	6.196595	0.8211267	727.63

Table 5.2: The three GLMs with estimated parameters; μ_i, σ_i and ν_i; and AIC values for zero-counts.

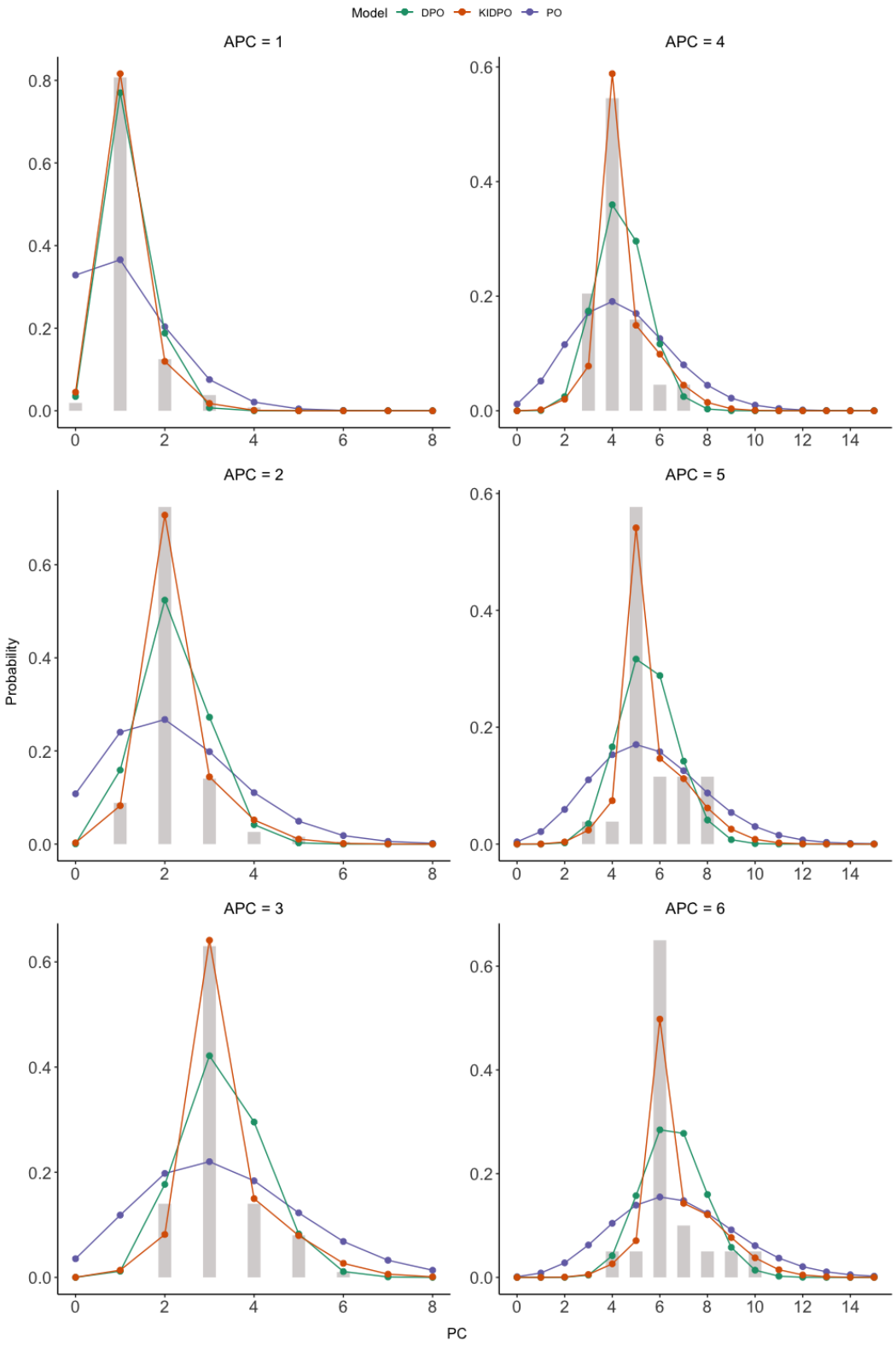


Figure 5.2: Fitted PMFs for the three GLMs compared to empirical data. The histograms shows the distribution of observed data.

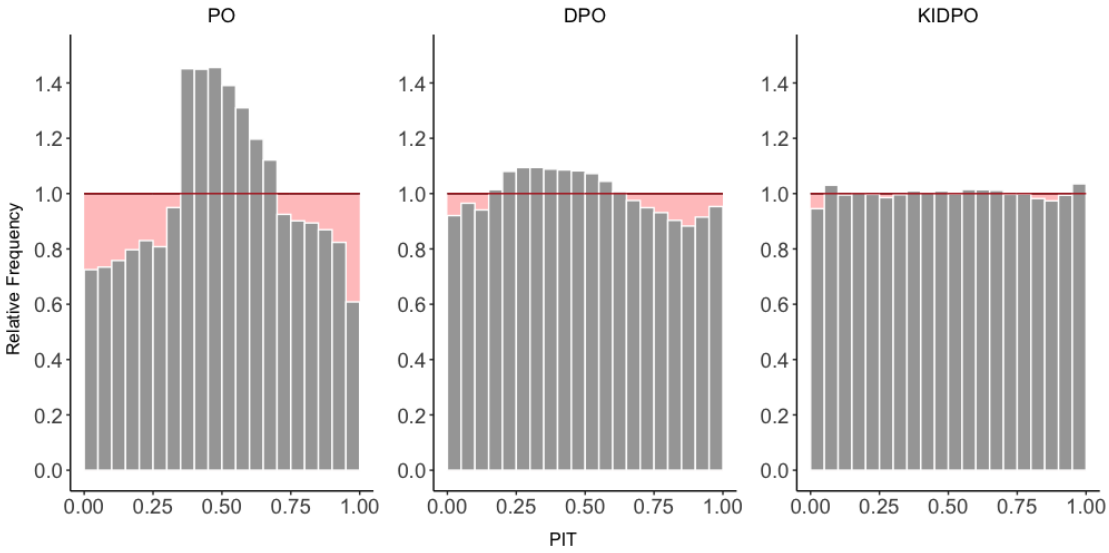


Figure 5.3: PIT histograms for the GLMs.

Model	AIC	Average RPS
PO	3071.241	0.1670
DPO	2395.596	0.1333
KIDPO	2223.909	0.1296

Table 5.3: Total AIC and average RPS for the GLMs.

5.1.2 Data Generating Process Model - *DGP-M*

We now consider the data generating process model *DGP-M* for door-level PC. Parameter estimates for the likelihood function are presented in Table 5.4. The total AIC and average RPS is listed in Table 5.5 along with that of *GLM*. Plots of the fitted PMF for selected APC values are presented in Figure 5.4. Figure 5.5 show plots of the *DGP-M* and *GLM* PMFs over empirical distributions and their PIT histograms are shown together in Figure 5.6.

Visually the *DGP-M* looks quite similar to the *GLM* for small values of APC. For larger values however they begin to differ substantially and we specifically note that the *DGP-M* fails to adequately capture the peak we see in the empirical data, this is due to the relatively strict assumptions we have made for *DGP-M*. Our suspicion from visual inspection is that *DGP-M* provides a slightly worse fit, this is belief is strengthened by consulting the AIC values. Though the AIC of *DGP-M* is of the same order of magnitude as that of the *GLM*-AIC, it is still markedly worse with $|\Delta_{AIC}| = 28.138$. RPS gives further evidence to *GLM* outperforming *DGP-M*, as it has a slightly higher average RPS. *DGP-M* is still an improvement on both *PO* and *DPO* though.

The PIT histogram of *DGP-M* looks very close to uniform for PIT values below 0.65 but then has a small dip and a peak. This indicates that when the model underestimates, it tends to underestimate by a larger amount then what would be ideal.

The model *DGP-M* provides a slightly worse fit than the *GLM*. It does however perform surprisingly well considering the simple, and straight forward assumed data generating process and we suspect an even better fit should be possible with some changes

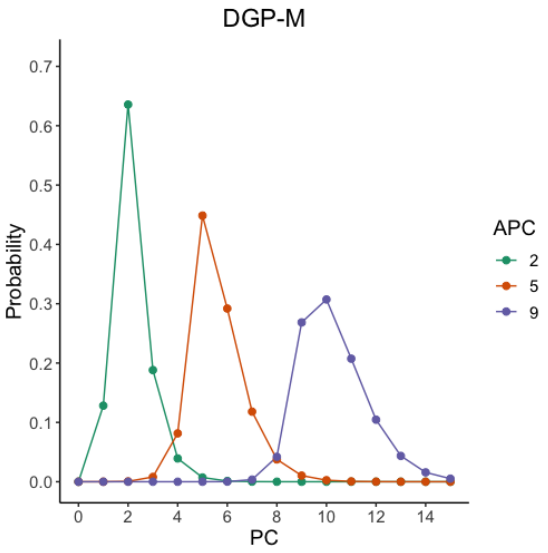


Figure 5.4: PMF of *DGP-M* shown for selected values of APC.

to the model assumptions.

Parameter	Estimate
p	0.8141423
μ	$\begin{cases} 0.09283607, \text{ if } PC \neq 0 \\ 0.00367642, \text{ if } PC = 0 \end{cases}$

Table 5.4: Parameter estimates of the likelihood function in *DGP-M*.

Model	AIC	Average RPS
GLM	2223.9	0.1296
DGP-M	2253.8	0.1304

Table 5.5: Total AIC and average RPS for *GLM* and *DGP-M*.

5.2 Results for Stop- and Journey-Level PC

We further consider *GLM* and *DGP-M* on stop and journey aggregates, PC_{stop} and $PC_{journey}$. Stop-level PIT histograms are shown in Figure 5.7 and journey-level in Figure 5.8. Figure 5.9 show PIT histograms for the random journeys. The average RPS of both models for stop-level aggregates are presented in Table 5.6, and average CRPS for journey-level aggregates in Table 5.7.

We first note that *GLM* outperforms *DGP-M* at stop-level in terms of RPS and at journey-level in terms of CRPS. This is what we would expect from *GLM* being the better model for door-level observations.

Next we observe that while both *GLM* and *DGP-M* had PIT histograms that closely mirror uniform distributions at the door-level, this uniformity does not extend to stop-

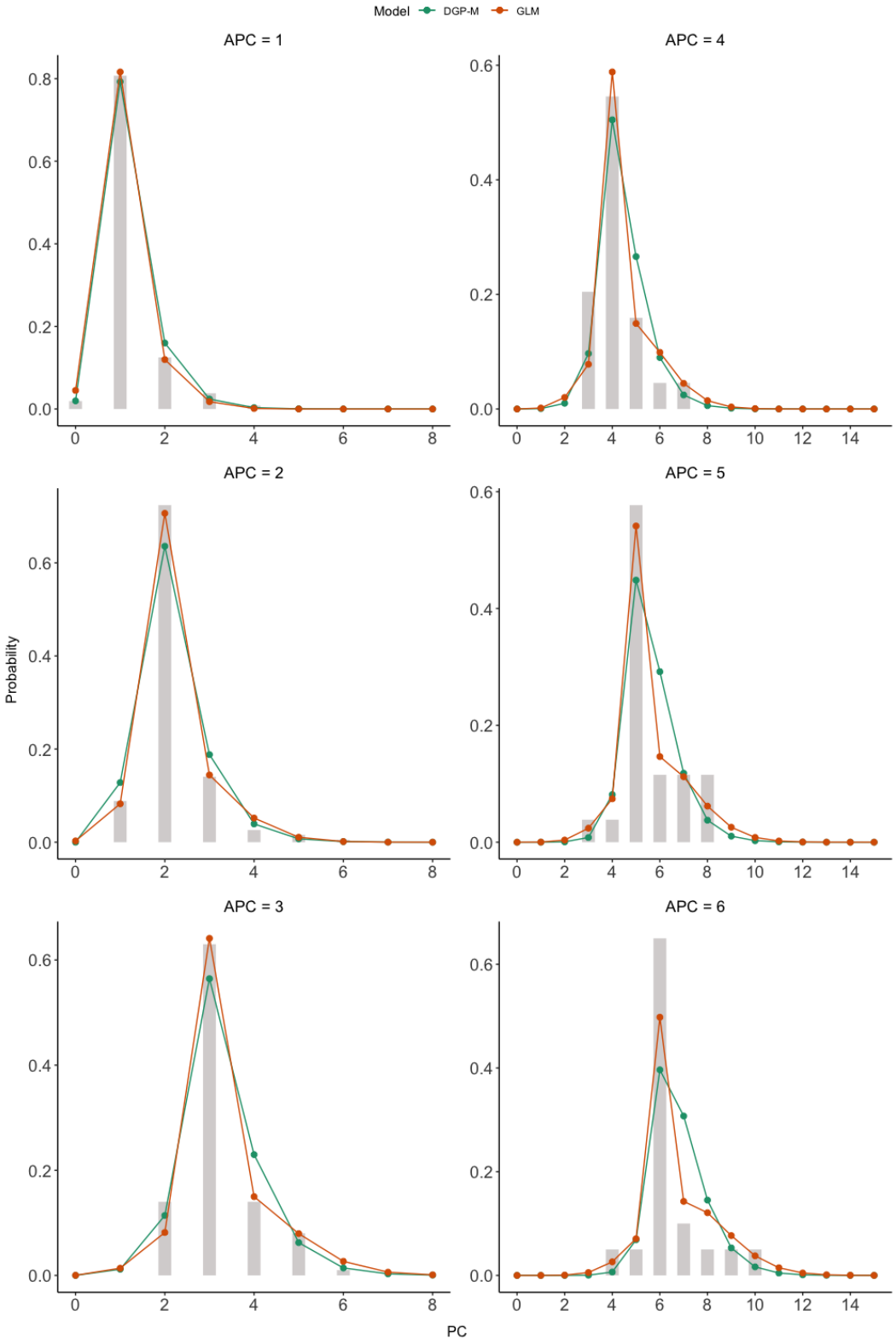


Figure 5.5: PMFs of *GLM* and *DGP-M*. The histogram shows the distribution of observed data.

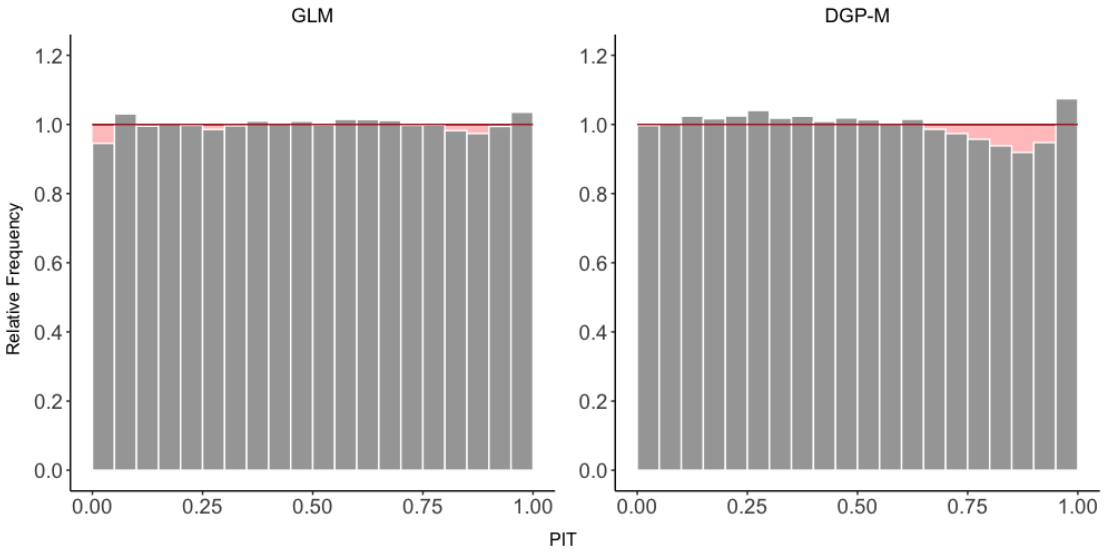


Figure 5.6: PIT histograms for *GLM* and *DGP-M* on door-level observations.

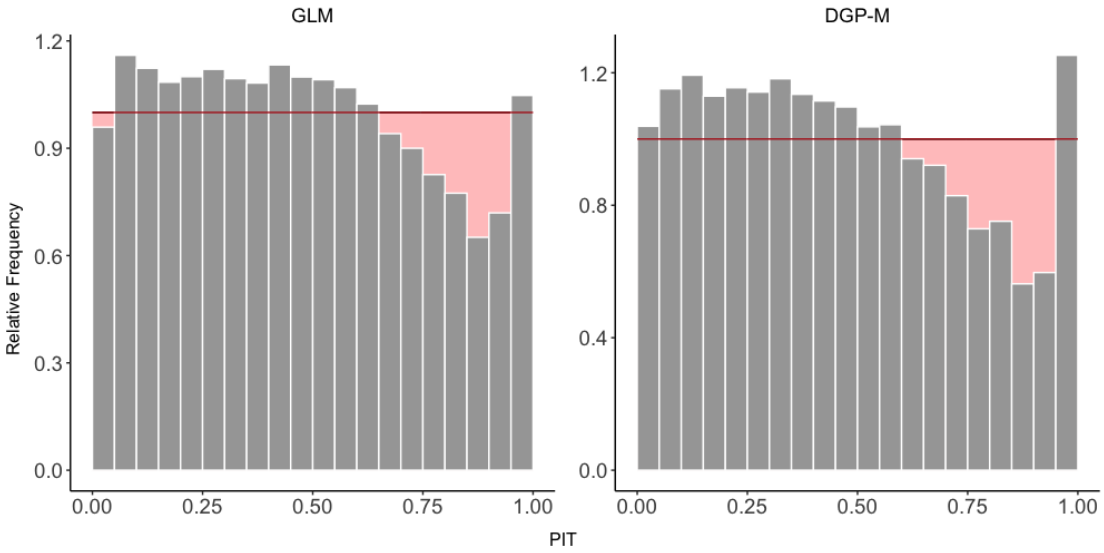


Figure 5.7: PIT histograms for *GLM* and *DGP-M* on stop-level aggregates.

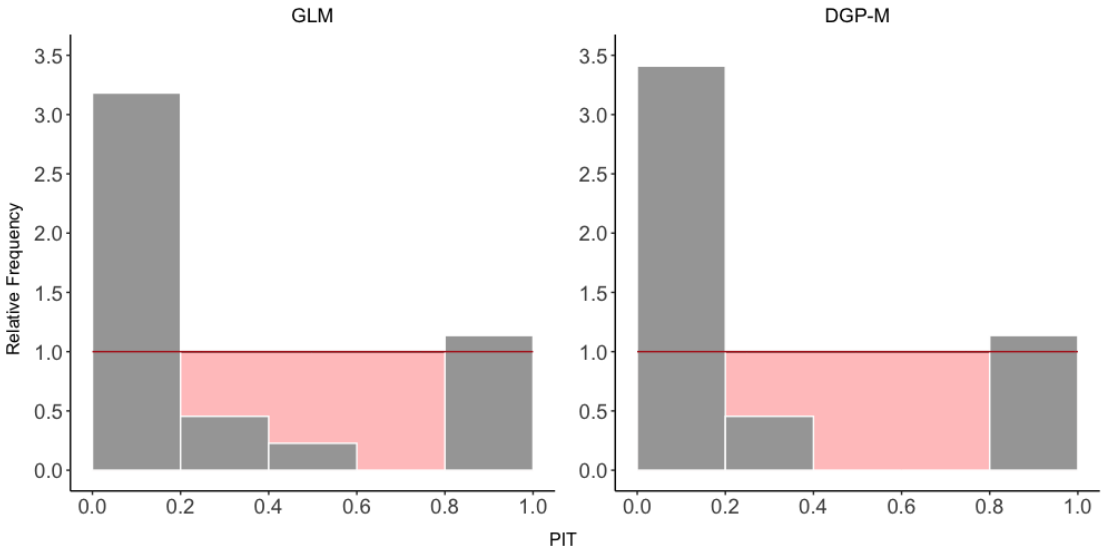


Figure 5.8: PIT histograms for *GLM* and *DGP-M* on journey-level aggregates.

Model	Average RPS
GLM	0.3945
DGP-M	0.3996

Table 5.6: Average RPS of *GLM* and *DGP-M* for prediction of stop-level aggregates.

Model	Average CRPS
GLM	10.9810
DGP-M	11.1352

Table 5.7: Average CRPS of *GLM* and *DGP-M* for prediction of journey-level aggregates.

and journey-level aggregates. Looking at the stop-level PIT histograms both models tend to overestimate, with PIT values below 0.5, too frequently. The peak at the right border of the PIT histograms are most likely due to outliers, this is then an indication that when the models don't overestimate, they tend to make substantial underestimations, with PIT values close or equal to 1.

For journey-level aggregates, we should interpret the PIT histograms with caution due to the limited number of data points, 22 (2.5). Nonetheless, the deviation from uniformity is striking, with the PIT histograms demonstrating a conspicuous U-shape. Both models exhibit a tendency to significantly overestimate the aggregates, while still sometimes producing extreme underestimates.

The discrepancy between the door-level PIT histograms and those at the stop- and journey-levels suggests the presence of correlations between observations at the same stops and journeys. This, in turn, indicates that the independence assumptions we made when applying the models to stop- and journey-level aggregates are not being met. This observation is further validated by the PIT histograms for the random journeys depicted in Figure 5.9, which do appear to more closely resemble a uniform distribution.

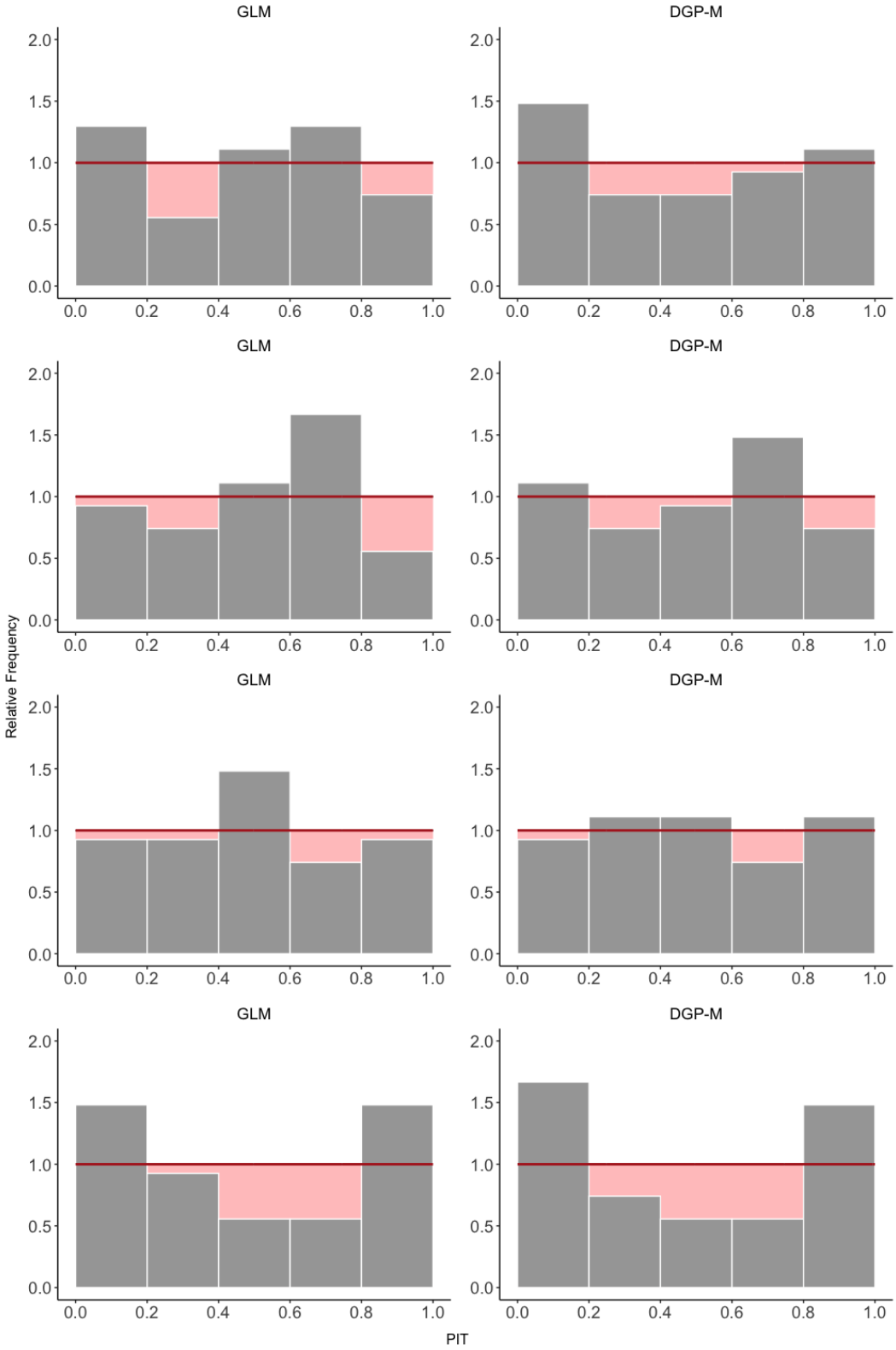


Figure 5.9: PIT histograms for *GLM* and *DGP-M* on aggregates of random groupings of 100 door-level observations. Four realizations shown.

6

Discussion and Final Remarks

In this thesis we have sought to develop door-level, probabilistic models for the passenger count, PC, using the automatic count, APC. Specifically a generalized linear model, GLM, was identified as desirable due to the complex data generating process, DGP, and the linear relation between APC and the expected value of PC. However, conventional GLMs fail to adequately model the specific characteristics of our data. The data is significantly underdispersed relative to the Poisson, and the double Poisson distribution is insufficient in modeling this underdispersion due to the high frequency of certain values. Although previous research has treated 0- and k-inflated distributions within the context of GLMs to rectify such issues, the current implementations are insufficient as they require the inflation point to be fixed.

In response to these challenges, we propose a GLM that utilizes the k-inflated double Poisson, KIDPO, in a way that allows the inflation point k to vary with the explanatory variable. This allows for the modeling of data where the marginal distributions are all KIDPO, but with variable inflation points k_i.

Initial evaluations of the proposed KIDPO-GLM indicate promising results for modeling this kind of data. It displays an excellent fit to the door-level observations, with the PIT histogram closely resembling a uniform distribution, suggesting a well-calibrated model. Additionally, it outperforms the other proposed model, *DGP-M*, which attempts to capture the true DGP, both in terms of AIC at the door-level and in terms of (C)RPS for stop- and journey-level aggregates. It is worth noting, though, that the *DGP-M* could likely be improved with a more rigorous analysis than undertaken in this thesis. However, such an improvement would necessitate a comprehensive understanding of the DGP and its potential sources of error.

The *DGP-M* model does offer some notable benefits. Firstly, it attempts to model the true DGP, offering a more representative depiction of the underlying phenomenon. Secondly, it provides a straightforward mechanism for incorporating prior knowledge on PC. Factors such as the time of day, bus stop location, and day of the week provide valuable insights into the amount of passengers that are expected. This information can be readily integrated into the model's prior assumptions, a feat that might pose a greater challenge with a GLM.

Applying the models to stop- and journey-level aggregates of PC, revealed an undeniable dependency among door-level observations. A portion of this dependency can likely be attributed to door-issues, as discussed in Section 2.1, on the specific $journey$

or at the specific $stop$. The door-issues tend to persist throughout a journey, resulting in extreme outliers relative to the rest of the data. Consequently, the models overestimate the mean due to the presence of outliers. When aggregating, this overestimation is compounded and the models tend to consistently overestimate the true PC aggregates when door-issues are absent, yet underestimate significantly in their presence.

To account for the intra-group dependency that exist within the $journey$ and $stop$ variables during the modeling process, we could incorporate certain random effects corresponding to these variables. However, it's crucial to note that these would be inherently unavailable for use in a practical scenario on other data. A strength of GLMs is that random effects can be included fairly easily. A random intercept for $journey$ on both or either of μ in the double Poisson-component and ν for the inflation, should be able account for the door-issues.

Incorporating random effects into the *DGP-M* model is less straightforward. One could include a random effect on the parameter p of the binomial component, or the parameter μ of the Poisson component. An additional possibility involves expanding the model to include a component that directly accounts for the door-issues. In general it does however provide more of a challenge than for GLMs, as we would need to better understand the source of dependency.

While we suspect that the door-issues are the primary source of dependency, other sources could potentially contribute. For instance, similar APC values might tend to occur together at stop- or journey-level, and the models might show varying performance across different ranges of APC. Additionally, inherent differences in the APC system performance may exist between different bus models due to variations in sensor placement or during different weather conditions. Employing the our proposed KIDPO-GLM can enable public transport agencies to identify the factors that impact the performance of their APC systems, and include effects for these.

For this thesis, we have chosen to omit data pertaining to alighting passengers due to substantial errors present in that data. Consequently, estimators for the aggregated number of passengers based on our models utilize a single count for each passenger. Incorporating alighting data could potentially enhance aggregate models as each passenger would essentially be counted twice. Furthermore, at the door-level, alighting passengers form an integral part of the true DGP. Thus it should be included as a variable in *DGP-M*, and could potentially improve the GLM as well.

Finally it should also be noted that the vast majority of the data set contains small values for APC and that there are very few data points in the upper range of APC. This imbalance makes it challenging to evaluate the performance of the model on larger counts. However as discussed in Section 2.3, these larger counts logically represent a relatively larger fraction of total passengers compared to their proportion of data points. Thus, any useful model should perform adequately also in this range. Consequently efforts should be made to gather more data also in the upper range of APC, in order to improve model fit in these ranges.

Bibliography

Abramowitz, M., 1965. Handbook of mathematical functions with formulas. Graphs, and Mathematical Tables .

Anscombe, F.J., 1950. Sampling theory of the negative binomial and logarithmic series distributions. Biometrika 37, 358–382.

Arora, M., Chaganty, N.R., 2021. Em estimation for zero-and k-inflated poisson regression model. Computation 9, 94.

Arora, M., Rao Chaganty, N., Sellers, K.F., 2021. A flexible regression model for zero-and k-inflated count data. Journal of Statistical Computation and Simulation 91, 1815–1845.

AtB, 2016. Framtidig rutestruktur med superbuss i Stor-Trondheim 2019-2029: Informasjonsteknologi og systemer. https://www.atb.no/getfile.php/132272-1509446090/Rapporter/AtB_Framtidig_rutestruktur_2019-2029_Informasjonsteknologi_systemer_13.05.16.pdf Accessed 13/05/2023.

Berrebi, S.J., Joshi, S., Watkins, K.E., 2022. Cross-checking automated passenger counts for ridership analysis. Journal of Public Transportation 24, 100008.

Boyle, D.K., 2008. Passenger counting systems. 77, Transportation Research Board.

Cameron, A., Trivedi, P., 2013. Regression Analysis of Count Data. Econometric Society Monographs, Cambridge University Press.

Casella, G., 1985. An introduction to empirical bayes data analysis. The American Statistician 39, 83–87.

Casella, G., Berger, R.L., 2002. Statistical Inference. 2nd ed., Duxbury press.

DILAX, . Automatic passenger counting (APC). https://www.dilax.com/en/products/automatic-passenger-counting Accessed 13/05/2023.

Dunn, P.K., Smyth, G.K., 1996. Randomized quantile residuals. Journal of Computational and graphical statistics 5, 236–244.

Efron, B., 1986. Double exponential families and their use in generalized linear regression. Journal of the American Statistical Association 81, 709–721.

Fahrmeir, L., Kneib, T., Lang, S., Marx, B., 2013. Regression: Models, Methods and Applications. Springer Berlin Heidelberg.

Givens, G.H., Hoeting, J.A., 2012. Computational statistics. volume 703. John Wiley & Sons.

Halyal, S., Mulangi, R.H., Harsha, M., 2022. Forecasting public transit passenger demand: With neural networks using apc data. Case Studies on Transport Policy 10, 965–975.

Hilbe, J., 2014. Modeling Count Data. Cambridge books online, Cambridge University Press.

Jordan, A., Krüger, F., Lerch, S., 2019. Evaluating probabilistic forecasts with scoringRules. Journal of Statistical Software 90, 1–37. doi:10.18637/jss.v090.i12.

Kimpel, T.J., Strathman, J.G., Griffin, D., Callas, S., Gerhart, R.L., 2003. Automatic passenger counter evaluation: Implications for national transit database reporting. Transportation research record 1835, 93–100.

Laboratory, N.R.A., 2015. verification: Weather Forecast Verification Utilities. URL: https://CRAN.R-project.org/package=verification. r package version 1.42.

Mccarthy, C., Moser, I., Jayaraman, P.P., Ghaderi, H., Tan, A.M., Yavari, A., Mehmood, U., Simmons, M., Weizman, Y., Georgakopoulos, D., et al., 2021. A field study of internet of things-based solutions for automatic passenger counting. IEEE Open Journal of Intelligent Transportation Systems 2, 384–401.

Mohammadpour, S., Stasinopoulos, M., 2018. gamlss.countKinf: Generating and Fitting K-Inflated 'discrete gamlss.family' Distributions. URL: https://CRAN.R-project.org/package=gamlss.countKinf. r package version 3.5.1.

Nagaraj, N., Gururaj, H.L., Swathi, B.H., Hu, Y.C., 2022. Passenger flow prediction in bus transportation system using deep learning. Multimedia tools and applications 81, 12519–12542.

Payandeh Najafabadi, A.T., Mohammadpour, S., 2018. A k-inflated negative binomial mixture regression model: application to rate–making systems. Asia-Pacific Journal of Risk and Insurance 12, 20170014.

R Core Team, 2021. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria. URL: https://www.R-project.org/.

Rigby, R.A., Stasinopoulos, D.M., 2005. Generalized additive models for location, scale and shape,(with discussion). Applied Statistics 54, 507–554.

Rigby, R.A., Stasinopoulos, M.D., Heller, G.Z., De Bastiani, F., 2019. Distributions for modeling location, scale, and shape: Using GAMLSS in R. CRC press.

Stasinopoulos, M., Rigby, R., 2022. gamlss.dist: Distributions for Generalized Additive Models for Location Scale and Shape. URL: https://CRAN.R-project.org/package=gamlss.dist. r package version 6.0-5.

Stasinopoulos, M.D., Rigby, R.A., Heller, G.Z., Voudouris, V., De Bastiani, F., 2017. Flexible regression and smoothing: using GAMLSS in R. CRC Press.

Strathman, J.G., 1989. An evaluation of automatic passenger counters: validation, sampling, and statistical inference. Center for Urban Studies Publications and Reports 109.

Transportøkonomisk institutt, 2022. Apt-r: Analysis of public transport data: Building an open-source library in r. https://www.toi.no/its/prosjekt-apt-r/ Accessed 13/05/2023.

Wickham, H., 2016. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York. URL: https://ggplot2.tidyverse.org.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L.D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T.L., Miller, E., Bache, S.M., Müller, K., Ooms, J., Robinson, D., Seidel, D.P., Spinu, V., Takahashi, K., Vaughan, D., Wilke, C., Woo, K., Yutani, H., 2019. Welcome to the tidyverse. Journal of Open Source Software 4, 1686. doi:10.21105/joss.01686.

Wilks, D.S., 2011. Statistical methods in the atmospheric sciences. volume 100. Academic press.

Zou, Y., Geedipally, S.R., Lord, D., 2013. Evaluating the double poisson generalized linear model. Accident Analysis & Prevention 59, 497–505.

Appendix

A Expectation And Variance of the K-inflated Double Poisson Distribution

The k-inflated double Poisson PMF is stated in (3.7), and using the accurate approximations for the expectation and variance of the double Poisson (3.6) we can derive expressions for the approximate expectation and variance of the k-inflated double Poisson distribution.

First we establish that for a double Poisson distributed random variable X, with PMF $g(x)$ (3.5):

$$E[X] = \sum_X x \cdot g(x)$$

and

$$\begin{aligned} V[X] &= E[X^2] - (E[X])^2 \\ &= \sum_X [x^2 \cdot g(x)] - (\sum_X x \cdot g(x))^2 \\ \sum_X [x^2 \cdot g(x)] &= V[X] + (\sum_X x \cdot g(x))^2. \end{aligned}$$

It follows from (3.6) that

$$\sum_X x \cdot g(x) \approx \mu, \tag{6.1}$$

and

$$\sum_X [x^2 \cdot g(x)] \approx \mu\sigma + \mu^2. \tag{6.2}$$

The PMF of a k-inflated double Poisson distributed random variable Y is

$$f(y \mid \nu, \mu, \sigma) = \begin{cases} \nu + (1-\nu)g(y \mid \mu, \sigma), & \text{if } y = k \\ (1-\nu)g(y \mid \mu, \sigma), & \text{if } y \neq k \end{cases} \tag{6.3}$$

where $g(y)$ is the double Poisson-PMF (3.5). The expected value of Y is derived as follows:

$$\begin{aligned} E[Y] &= \sum_Y y \cdot f(y) \\ &= 0(1-\nu)g(0) + 1(1-\nu)g(1) + \ldots + k\nu + k(1-\nu)g(k) + \ldots \\ &= k\nu + (1-\nu) \cdot \sum_Y y \cdot g(y) \end{aligned}$$

and it follows from (6.1) that

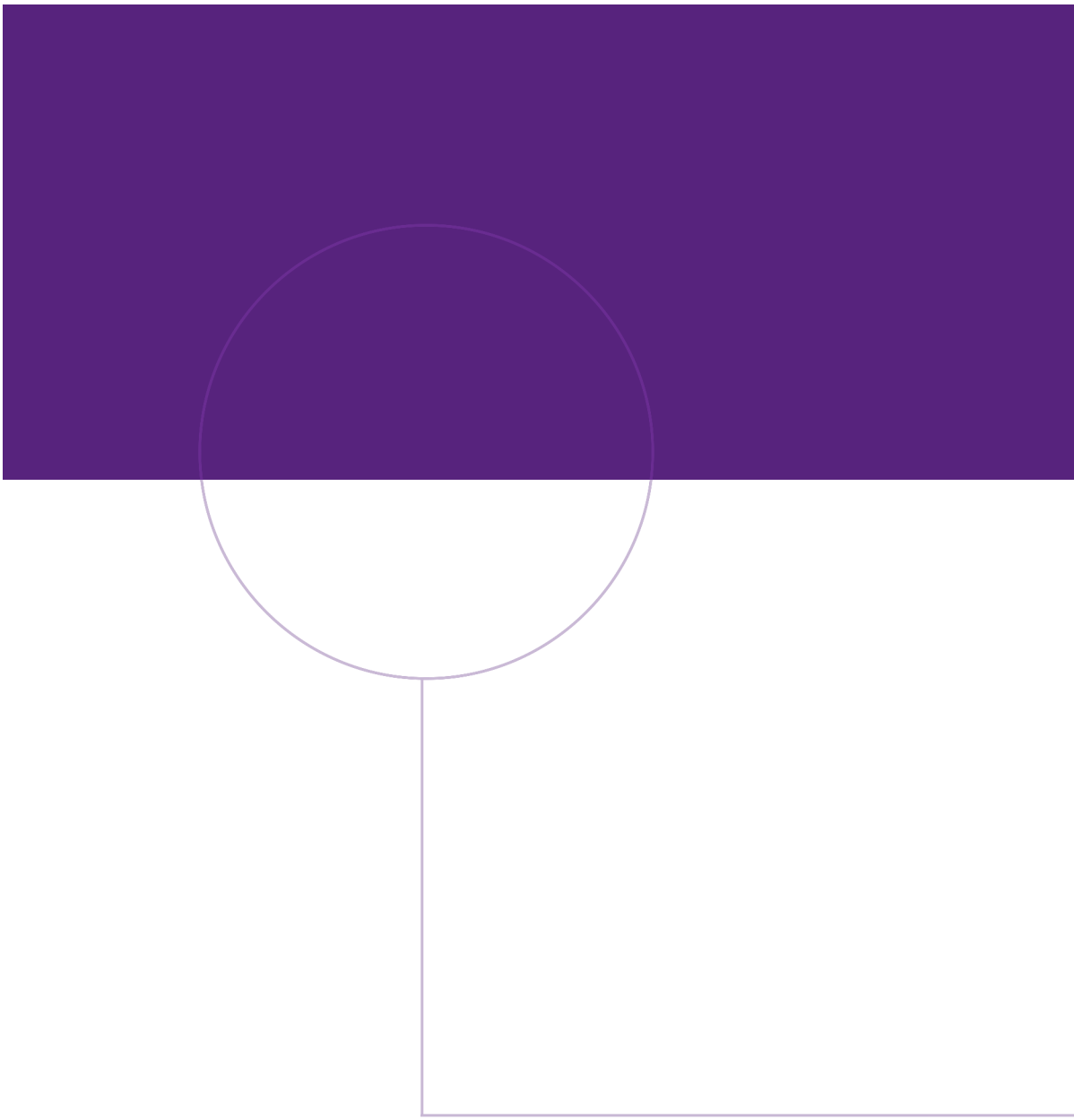
$$E[Y] \approx k\nu + (1-\nu) \cdot \mu. \tag{6.4}$$

The variance is derived similarly

$$\begin{aligned} E[Y^2] &= \sum_Y y^2 \cdot f(y) \\ &= 0^2(1-\nu)g(0) + 1^2(1-\nu)g(1) + \ldots + k^2\nu + k(1-\nu)g(k) + \ldots \\ &= k^2\nu + (1-\nu) \cdot \sum_Y y^2 \cdot g(y) \\ V[Y] &= E[Y^2] - (E[Y])^2 \\ &= k^2\nu + (1-\nu) \cdot \sum_Y y^2 \cdot g(y) - (E[Y])^2 \end{aligned}$$

and it follows from (6.1) and (6.2) that

$$V[Y] \approx \nu(1-\nu)(k-\mu)^2 + (1-\nu)\mu\sigma. \tag{6.5}$$



NTNU

Norwegian University of
Science and Technology