To Be Invariant or to Be Time-Varying: The Issue of Including Event-Related Covariates in the Zero-Truncated Poisson Regression Model

Annemarie Timmers (6238106)

Bachelorproject sociologie (201100018)

Under the supervision of Dr. M. Cruyff

Department of Methodology and Statistics

Utrecht University

June 2020

Abstract

The zero-truncated Poisson regression model is used to estimate the total size of populations that cannot be counted directly. In this model, covariates are used to account for variation within the population and produce an accurate estimate of the total population size. This study investigates the inclusion of event-related covariates in the model. These are covariates that represent a property of the capture itself that is only observed with the occurrence of a capture and whose value may vary over time. Consequently, multiple values may apply to one individual. However, since this information is mostly unobserved, including event-related covariates in the model is complicated. A simulation study is performed to investigate how event-related covariates should be included in the zero-truncated Poisson regression model. Three simulations are conducted in which an event-related covariate with two categories is respectively considered invariant and time-varying in the first two simulations, and a mixture of both in the third. Four methods are evaluated in the simulations: no covariate, one dummy, two dummies, and two count variables. This study shows that event-related covariates should not be included in the model when their value is assumed to be at least partially time-varying. Including event-related covariates regardless ensues misspecification of the model that results in biased estimates.

Keywords: capture-recapture, population size estimation, truncated Poisson regression, time-varying covariates

To Be Invariant or to Be Time-Varying: The Issue of Including Event-Related Covariates in the Zero-Truncated Poisson Regression Model

Capture-recapture methods are used to estimate the size of populations that cannot be counted directly. The zero-truncated Poisson regression model can be used when count data is available in a single registration file with the number of captures of the observed population members recorded during a fixed observation period. Individuals that were never captured have a count of zero and are therefore not on the list. Under the model, a Poisson parameter is estimated that defines the probabilities of a population member to be captured once, twice, and so on. Based on this Poisson parameter, the probability of a zero count can be estimated, so the number of remaining population members can be determined and added to the population members that were already on the list, resulting in an estimate of the total population size.

One important assumption of the Poisson distribution is the homogeneity of the Poisson parameter. This homogeneity means that the chances to be captured once, twice, and so on are assumed to be the same for all population members. However, it is often unrealistic to expect that this assumption holds ("Student", 1919). For example, men and women may have different Poisson parameters or the Poisson parameter may change with age (see Kromhout, Wubs, & Beenakkers, 2008; Cruyff & Van Der Heijden, 2008; Snippe & Mennes, 2018). Violation of the homogeneity assumption can result in an underestimation of the total population size. As such, it is essential to deal with this heterogeneity. This can be done by specifying the Poisson parameter as a function of covariates in Poisson regression, which prompts the estimation of multiple Poisson parameters that can vary based on an individual's scores on the covariates (Cameron & Trivedi, 1998). Subsequently,

one would want to add covariates to the model that account for the heterogeneity in the population.

There are multiple types of covariates. For instance, covariates can be classified as continuous or categorical, intrinsic or extrinsic, invariant or time-varying, and internal or external. Since the type of covariate often dictates how they ought to be included in the model, it is beneficial to classify them based on their properties. In this, time-varying covariates pose a special challenge in Poisson regression, because the event counts are measured over the time in the observation period, but summed overall (Rostgaard, 2008). As such, the time component is lost in the model. The inclusion of time-varying covariates is then complicated, as their value is only observed with the occurrence of a capture. So, while these covariates may take on different values for the same individual during the observation, their variation over time is undetermined, and we lack the information to accurately predict the rest (Bonner, Thomson, & Schwarz, 2009). Further specifying the classification of time-varying covariates may provide directions as to how they can be included in the Poisson regression model. For example, since the change over time for age is predetermined, age can be observed once and fixed for the remaining part of the observation period.

However, within the criminological context of population estimation, we have come across a type of covariate that is potentially time-varying, and that stands out because of its ambiguous nature. Apart from the difficulty of a time-varying covariate as described above, the inclusion of this type of covariate is complicated, because it represents a property of the event count under study. Therefore, it cannot exclusively be regarded as an

individual characteristic when its value changes over time. As such, this could be regarded as an event-related covariate.

We consider the police district in which the capture took place as an example. This covariate was taken into account by Van Der Heijden et al. (2003b) in estimating the number of drunk drivers in the Netherlands. When capture priorities differ across police districts, their inclusion in the model is beneficial to explain a part of the population heterogeneity as individuals could move across police districts and consequently be subject to different amounts of tests (Hoogteijling, 2002). Then, the value of the police district is mostly invisible, as it is only observed with the occurrence of a capture. Moreover, this event-related covariate can no longer be considered as a property of the individual when different capture priorities of multiple police districts apply to the same individual during the observation period. Hence, the inclusion of event-related covariates in the Poisson regression model is complicated.

We may look at various types of covariates that are distinguished in the literature and attempt to classify this event-related covariate accordingly to gain insights into its inclusion. First, Catchpole, Morgan, and Tavecchia (2008) differentiated between extrinsic and intrinsic covariates for animal capture-recapture data. In this, extrinsic covariates relate to all population members, while intrinsic covariates depend on individual characteristics. Furthermore, since captures are recorded over an observation period, the value of the covariates may change over time. Therefore, both extrinsic and intrinsic covariates can be classified as invariant when their value is constant, or time-varying when their value changes over time (Bonner, Morgan, & King, 2010).

Multistate capture-recapture models distinguish between static and dynamic states as covariates. Even though this contrast is similar to that of the previous paragraph, states are categorical variables, while invariant and time-varying covariates can also refer to continuous variables. For static states, the value of the covariate does not change over time, like gender. In contrast, the value of dynamic states varies over time. This variation over time can further be divided into a change that occurs in a deterministic or probabilistic way. Deterministic change occurs at a constant rate over time, like age. For a probabilistic change, the amount of variation over time is not readily defined (Kendall, Conn, & Hines, 2006).

In survival analysis, the distinction between invariant/static and time-varying/dynamic covariates is found as well. Kalbfleisch and Prentice (1980) additionally defined two classes of time-varying covariates: external and internal (Austin, Latouche, & Fine, 2020; Cortese & Andersen, 2010). When a covariate is external, it may influence survival, but is not affected by the occurrence of failure, e.g., death, over time. Internal covariates, however, are reciprocally related to the outcome: they affect survival but are also affected by the survival itself. Internal covariates mostly follow a stochastic process that is often only observed as long as the individual survives. Later, Lancaster (1990) proposed a similar distinction, defining a covariate exogenous when the process by which the variable comes about depends on the process generating the outcome variable, and endogenous otherwise. In this, exogenous and endogenous covariates compare to external and internal covariates, respectively (Vermunt, 1996).

Kalbfleisch and Prentice (1980) further divided external covariates into three

categories: fixed, defined, and ancillary. The value of fixed external covariates is measured before the study and held constant for its duration. Similar to deterministic dynamic states, the change in the value of defined external covariates is predetermined for the duration of the study. Ancillary external covariates resemble probabilistic dynamic states because their value changes stochastically, following a process that lies outside the scope of the study.

When applied to the police district in which the capture took place as an event-related covariate, it becomes clear that no ready-made solution can be found, as the classification of an event-related covariate depends on unknown information; whether its value is time-varying. On the one hand, when its value is invariant, the police districts function as an intrinsic covariate. That is, an individual stays in one district during the observation period and can only be captured there. Then, the capture priority of a police district is an individual characteristic. On the other hand, when its value is time-varying and individuals can move across multiple districts, the police districts function as an extrinsic covariate. In this case, the capture priority of a district can no longer be regarded as a property of the individual. Rather, it is a property of the police district itself, equal for all individuals residing in that district throughout a portion of the observation period. This is in contrast to marital status, for example. Whether an individual's marital status changes during the observation period does not affect its classification as an intrinsic covariate.

However, the ambiguity does not end there, as the classification of the police district in which the capture took place under the other categories also depends on whether its

value can change over time. For example, invariant police districts are fixed external/exogenous, because their value is constant and cannot be affected by the capture. Meanwhile, the variation over time for time-varying police districts is deemed probabilistic as individuals are not assigned to a certain police district and can move across districts in a fluctuating manner. Likewise, it is internal/endogenous, because the value of the police district is only observed when a capture takes place.

So, as the classification of event-related covariates is contingent on its largely unobserved variation over time, it remains unclear how we should include them in the model. To simplify the discussion, we consider multiple approaches to the case that there are two police districts. First, we could disregard the police district in which the capture took place as a covariate. Second, we could use a dummy variable to mark the police district in which most captures took place, which was the approach taken by Van Der Heijden et al. (2003b) in estimating the number of drunk drivers in the Netherlands. Third, we could add two dummy variables, one for each district, where a one means that the individual was captured at least once in the district the dummy represents and a zero that no captures have taken place in that district. Lastly, two count variables could be included in the model, with a separate variable for each district in which one count equals one capture in that district, two counts equal two captures, and so on.

The purpose of this study is to evaluate the performance of these four methods for including event-related covariates in the zero-truncated Poisson regression model. This will be done by comparing their estimates of the total population size and Poisson parameters for accuracy in a simulation study. Three simulations will be performed. In the first two

simulations, the police district in which the capture took place will be regarded as an invariant and time-varying covariate, respectively. The third simulation will combine the properties of the preceding simulations in considering a population in which the event-related covariate manifests itself as both invariant and time-varying.

The paper is structured as follows. Section 2 discusses the zero-truncated Poisson regression model and related matters. Hereafter, the structure of the simulation studies and their results are described in section 3. The paper ends in section 4 with a conclusion based on the results and a discussion of its implications, a critical evaluation of the current study, and topics for further research.

2 The Model

In this section, the zero-truncated Poisson regression model is examined and the Horvitz-Thompson estimator for the total population size is described. Since the zero-truncated Poisson regression model is a modification of the Poisson regression model, which is derived from the Poisson probability distribution, these models and their features are also reviewed.

2.1 The Poisson Distribution and Zero-Truncation

The Poisson distribution is a theoretical probability distribution that describes the occurrence of events within a fixed observation period that are generated by a Poisson process, which is a series of random events occurring in time (Kingman, 1993). The probability distribution for an event to occur a number of times by a Poisson process is

denoted by

$$P(Y = y \mid \lambda) = \frac{\exp(-\lambda)\lambda^y}{y!}, \qquad y = 0, 1, 2, \ldots, \tag{1}$$

where the random variable Y for $y = 0, 1, 2, \ldots$ represents the number of observed events during the observation period, and $\lambda(\lambda > 0)$ is the Poisson parameter that jointly describes the mean and the variance of the distribution. This is always a number above zero, since it is assumed that there is a non-zero possibility for every population member to be captured.

When measured in discrete time, the Poisson parameter is a summation of the rates for all time units in the observation period T, given by

$$\lambda = \sum_{t=1}^{T} r_t, \tag{2}$$

where t denotes a discrete time unit, and r_t the number of events per time unit t. Generally, it is presumed that the rate is constant for the duration of the observation period. While we discuss this presumption in more detail in Section 2.3, it is important to note that this is not a requirement for the Poisson distribution to accurately describe an event count.

The number of captured drunk drivers can be described by a Poisson distribution as well. The Poisson parameter for a drunk driver would then correspond to the sum of the rates over the days in the observation period. The higher the value of the rate, the higher the chance for a drunk driver to be captured on a day in the observation period.

Since the registration file of the police does not contain the drunk drivers that were never captured, their number needs to be estimated from the observed events $y > 0$. This

can be incorporated in the expression of the Poisson distribution (1) by specifying the condition that $y > 0$, resulting in the zero-truncated Poisson distribution.

$$P(Y = y \mid y > 0, \lambda) = \frac{P(y \mid \lambda)}{P(y > 0 \mid \lambda)} = \frac{\exp(-\lambda)\lambda^y}{y!(1 - \exp(-\lambda))}, \qquad y = 1, 2, \ldots, \tag{3}$$

where $P(y > 0|\lambda) = 1 - \exp(-\lambda)$.

Following Van Der Heijden et al. (2003a), who derived the Horvitz-Thompson point estimate for the zero-truncated Poisson regression model, the estimate of the total population size N can be calculated as

$$\hat{N} = n\frac{1}{1 - P(Y = 0 \mid \lambda)} \tag{4}$$

where n are the population members in the registration file, and $P(Y = 0 \mid \lambda) = \exp(-\lambda)$ denotes the probability of a zero count. So, the lower the Poisson parameter, the higher the probability of a zero count, the higher the estimate of the total population size.

2.2 The Homogeneity Assumption and Poisson Regression

For an event count to be accurately described by a Poisson distribution, it has to adhere to the assumption of homogeneity. This means that the event count follows a Poisson distribution with the same Poisson parameters for all population members. As mentioned in the introduction, this assumption is often violated ("Student", 1919). In that case, the variance exceeds the mean and thus breaches the Poisson property that the

Poisson parameter jointly describes the mean and the variance. This is called overdispersion and ensues misspecification of the model (Hinde & Demétrio, 1998).

Van Der Heijden et al. (2003a) have demonstrated that failing to account for heterogeneity in the model results in an underestimation of the total population size through a special case of Jensen's inequality. Figure 1 shows the difference in the expectation of a zero count between a model with a dichotomous covariate and an intercept-only model. Referring to Section 2.1, the expectation of a zero count can be expressed as $\hat{P}(Y = 0)$ as it is the estimate of $P(Y = 0)$. Similarly, as the expectation of a zero count is higher for the model with the covariate, its estimate of the total population size is also higher compared to the intercept-only model. In this figure, δ indicates the average difference between the Poisson parameters of the model with the covariate and the intercept-only model. The higher its value, the bigger the difference between the models in the expectation of the zero counts, and thus the estimate of the population size.

The misspecification of a zero-truncated Poisson distribution has more far-reaching consequences than that of a Poisson distribution that is not zero-truncated. In contrast to a fully observed distribution, the mean of a zero-truncated distribution has yet to be determined. Therefore, the misspecification of a zero-truncated distribution can result in a biased estimate of the mean (Cameron & Trivedi, 1998).

Violation of the homogeneity assumption can be adjusted with the use of covariates, as these allow for the estimation of multiple different Poisson parameters. The zero-truncated Poisson distribution (3) can account for heterogeneity with the inclusion of

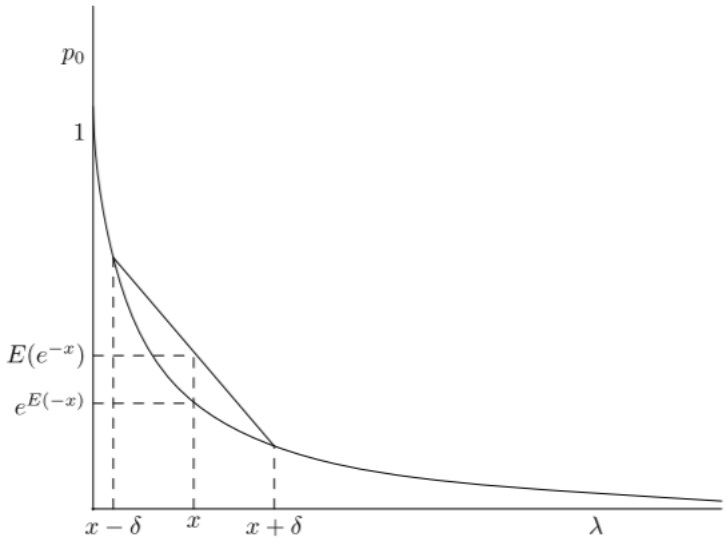

Figure 1. Illustration demonstrating that the model with a dichotomous covariate has a higher expectation of the zero count than the model without it. Reprinted from "Estimating the Size of a Criminal Population from Police Records Using the Truncated Poisson Regression Model", by P. G. Van Der Heijden, M. Cruyff and H. C. Van Houwelingen, 2003, *Statistica Neerlandica, 57*, p. 298.

covariates, given by

$$P(Y_i = y_i \mid y_i > 0, \lambda_i) = \frac{P(y_i \mid \lambda_i)}{P(y_i > 0 \mid \lambda_i)} = \frac{\exp(-\lambda_i)\lambda_i^{y_i}}{y_i!(1 - \exp(-\lambda_i))}, \qquad y_i = 1, 2, \ldots, \tag{5}$$

where the subscript i indicates that the Poisson parameter is not the same for all population members. Instead, the Poisson parameter of an individual λ_i can be derived through the *exponential mean function*

$$\lambda_i = \exp(\boldsymbol{x}_i'\boldsymbol{\beta}), \tag{6}$$

where $\boldsymbol{x}_i'$ is the value of the covariate specific to each population member, and $\boldsymbol{\beta}$ represents the corresponding parameter (Cameron & Trivedi, 1998). Hereby, the population is divided

into multiple homogenous subgroups as the Poisson parameters are still assumed to be the equal for individuals with identical scores on the covariates. With this adjustment, the zero-truncated Poisson regression model (5) can now also account for varying capture priorities of drunk drivers across police districts.

Heterogeneity can also be incorporated into the Horvitz-Thompson estimator of the total population size (4), denoted by

$$\hat{N} = \sum_{i=1}^{n} \frac{1}{1 - P(Y_i = 0 \mid \lambda_i)}, \tag{7}$$

where $P(Y_i = 0 \mid \lambda_i) = \exp(-\lambda_i)$.

The proportion of heterogeneity accounted for by the covariates is called observed heterogeneity. The remaining heterogeneity that is not captured in the model is called unobserved heterogeneity and still violates the homogeneity assumption, which may result in the underestimation of the total population size (Yamaguchi, 1986).

2.3 Independence of Rate and Capture

Another assumption of the Poisson model is the independence of rate and capture. This assumption implies that the rate of a population member should not change as a result of a capture (Nelder & Wedderburn, 1972). A violation of this assumption is known as *contagion*. Positive contagion takes place when the occurrence of a capture causes the rate to increase. Negative contagion occurs when the rate decreases as a consequence of a capture taking place.

Often, this assumption is interpreted in the sense that the rate should be constant

during the observation period (Lovett & Flowerdew, 1989). However, by Raikov's theorem, when the sum of two randomly distributed variables follows a Poisson distribution, then so do each of the separate random variables (Raikov, 1938). This theorem also holds in reverse, see for example Upton and Cook (1996) or Johnson, Kemp, and Kotz (2005). Then, when the random variable X_1 follows a Poisson distribution with the Poisson parameter λ_1, and the random variable X_2 is also a realization of the Poisson distribution with the Poisson parameter λ_2, the sum of X_1 and X_2 is described by a Poisson distribution with Poisson parameter $\lambda_1 + \lambda_2$.

Consequently, a Poisson distribution that describes an event count for an observation period may be split up in multiple subperiods, in which the event count is still characterized by a Poisson distribution. Therefore, the rate may vary over the subperiods, as long as a capture does not induce the change.

Applied to the example of the police districts, a capture should not cause a change in the rate. A drunk driver should not drive around drunk more often as a result of not being captured, which is positive contagion. Equally, negative contagion should not occur when a drunk driver stops drinking after a capture during the observation period. However, as long as the rate does not change because a capture has taken place, rates may also vary between police districts. Consequently, it is permitted to sum the rates of the police districts and calculate the Poisson parameter of an individual, employing

$$\lambda_i = \sum_{t=1}^{T} r_{ijt}, \tag{8}$$

where r_{ijt} denotes the average number of captures r in police district j where individual i

resides at time t.

3 Simulation Studies

We performed three simulations to evaluate the performance of four methods for including an event-related covariate in the zero-truncated Poisson regression model. Herein, the example of the police district in which the capture took place was continued. For simplification, only two police districts were considered. In the first simulation, the districts represented an invariant event-related covariate. Therefore, a group was assigned to each police district that was assumed to stay in that district for the whole observation period. This constraint was lifted in the second simulation where the police districts were considered as a time-varying event-related covariate. Consequently, one group was considered that divided its time between the two police districts. In the third simulation, the preceding simulations were combined and we examined a population with a mixed presence of the police districts as invariant and time-varying covariates. A group was assigned to each district that stayed there permanently, while a third group divided its time between the two districts.

The two police districts, characterized by the letters R and A, were incorporated as an event-related covariate into the zero-truncated Poisson regression model with the use of four methods, being:

1. No covariates: the police districts were not taken into account with a variable, resulting in an intercept-only model.

2. One dummy: the police districts were included in the model with one dummy, where a one implied that the number of captures of a population member was equal or

greater in district A than the number of captures in district R, and a zero meant that the number of captures was greater in district R.

3. Two dummies: the police districts were incorporated into the model with two dummy variables, each representing a police district. A one equaled that at least one capture of a population member took place in the district of the dummy, and a zero indicated that no captures took place in that district.

4. Two count variables: the police districts were added to the model with two count variables, one for each district. One count equaled one capture of a population member in the police district the variable represented, a two equaled two captures in that district, and so on.

In every simulation, a population of size $N = 10,000$ was randomly drawn from a Poisson distribution in accordance to the specified conditions. When necessary, additional variables were created to resemble one of the four methods. For all conditions, the population was generated 500 times, saving a zero-truncated sample that encompassed roughly 20-25% of the total population. The zero-truncated Poisson regression model in which one of the four methods was used to include the event-related covariate was fitted to all $13,500$ samples, estimating the total population size, its accompanying 95% confidence interval, and the Poisson parameters for the intercept-only and one dummy models.

3.1 Simulation 1

In the first simulation, the event-related covariate was considered invariant. We varied the group sizes and district-specific rates in nine conditions, depicted in Table 1. The two dummy method was omitted in the first simulation, because it would be a copy of

the one dummy method as the population members could only be captured in one district.

Table 1

Conditions in Simulation 1

Condition	n_R	n_A	r_R	r_A	T	λ_R	λ_A
1	5,000	5,000	0.0025	0.0025	100	0.25	0.25
2	2,500	7,500	0.0025	0.0025	100	0.25	0.25
3	7,500	2,500	0.0025	0.0025	100	0.25	0.25
4	5,000	5,000	0.0035	0.0015	100	0.35	0.15
5	2,500	7,500	0.0035	0.0015	100	0.35	0.15
6	7,500	2,500	0.0035	0.0015	100	0.35	0.15
7	5,000	5,000	0.0015	0.0035	100	0.15	0.35
8	2,500	7,500	0.0015	0.0035	100	0.15	0.35
9	7,500	2,500	0.0015	0.0035	100	0.15	0.35

Note. $N = 10,000$.

3.1.1 Point and interval estimates of N. The boxplots in Figure 2 show the performance of the three models through the distribution of the population size estimates in the respective conditions. Table 2 reports the corresponding coverage probabilities.

The intercept-only model estimated the total population size accurately when the rates were homogeneous. However, as soon as heterogeneity was introduced in condition 4, this model underestimated the total population size. The coverage probabilities reflect this; while the coverage probabilities were sufficient for the homogeneous rates, they dropped below the nominal level of 95% in the presence of heterogeneity.

The estimates of the one dummy model were accurate in the first simulation. Regardless of condition, this model produced valid estimates of the total population size. Besides, the coverage probabilities were stable at approximately 95%.

The count model overestimated the total population size in all conditions. Its overestimation was most extreme when the rates were homogeneous. Consequently, the

coverage probabilities of the count model were subnominal. Lastly, for all methods, group size did not affect the estimates of the total population size.

Table 2

95% Coverage Probabilities per Condition for All Methods in Simulation 1

	Coverage probabilities		
Condition	Intercept-only	One dummy	Count
1	0.956	0.958	0.000
2	0.946	0.950	0.000
3	0.934	0.940	0.000
4	0.178	0.948	0.000
5	0.316	0.946	0.004
6	0.502	0.966	0.000
7	0.226	0.970	0.000
8	0.558	0.956	0.000
9	0.332	0.940	0.002

3.1.2 Point estimates of λ. The boxplots in Figure 3 display the estimates of the Poisson parameters for the intercept-only (0) and one dummy models (R and A). The colored lines in the figures indicate the true values of the Poisson parameters, the boxplots represent what was estimated by the zero-truncated Poisson regression model.

Similar to the estimates of the total population size, the estimates of the intercept-only model were unbiased when the rates were homogeneous. For heterogeneous rates, the lower the estimate of the Poisson parameter, the more the total population size was underestimated. In contrast, the one dummy model produced unbiased estimates of the Poisson parameters in all conditions. As the group size increased, the estimates of the Poisson parameters were spread further apart. This can be seen on the right side of Figure

3, where the biggest colored boxplot corresponds to the police district of which the group size was set to be the largest.

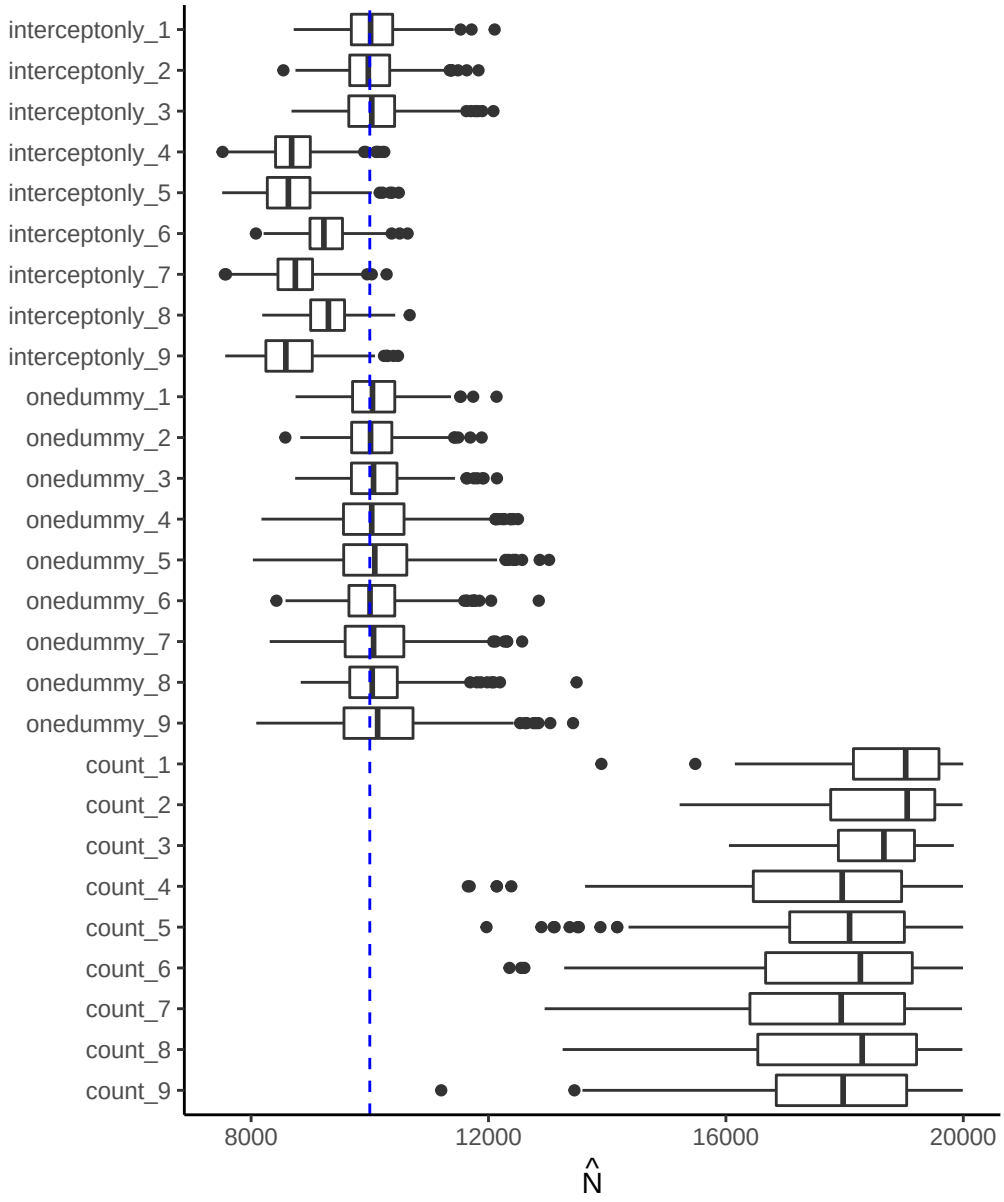

Figure 2. Boxplots displaying the distribution of the population size estimates for simulation 1.

Note. Dashed line depicts true total population size.

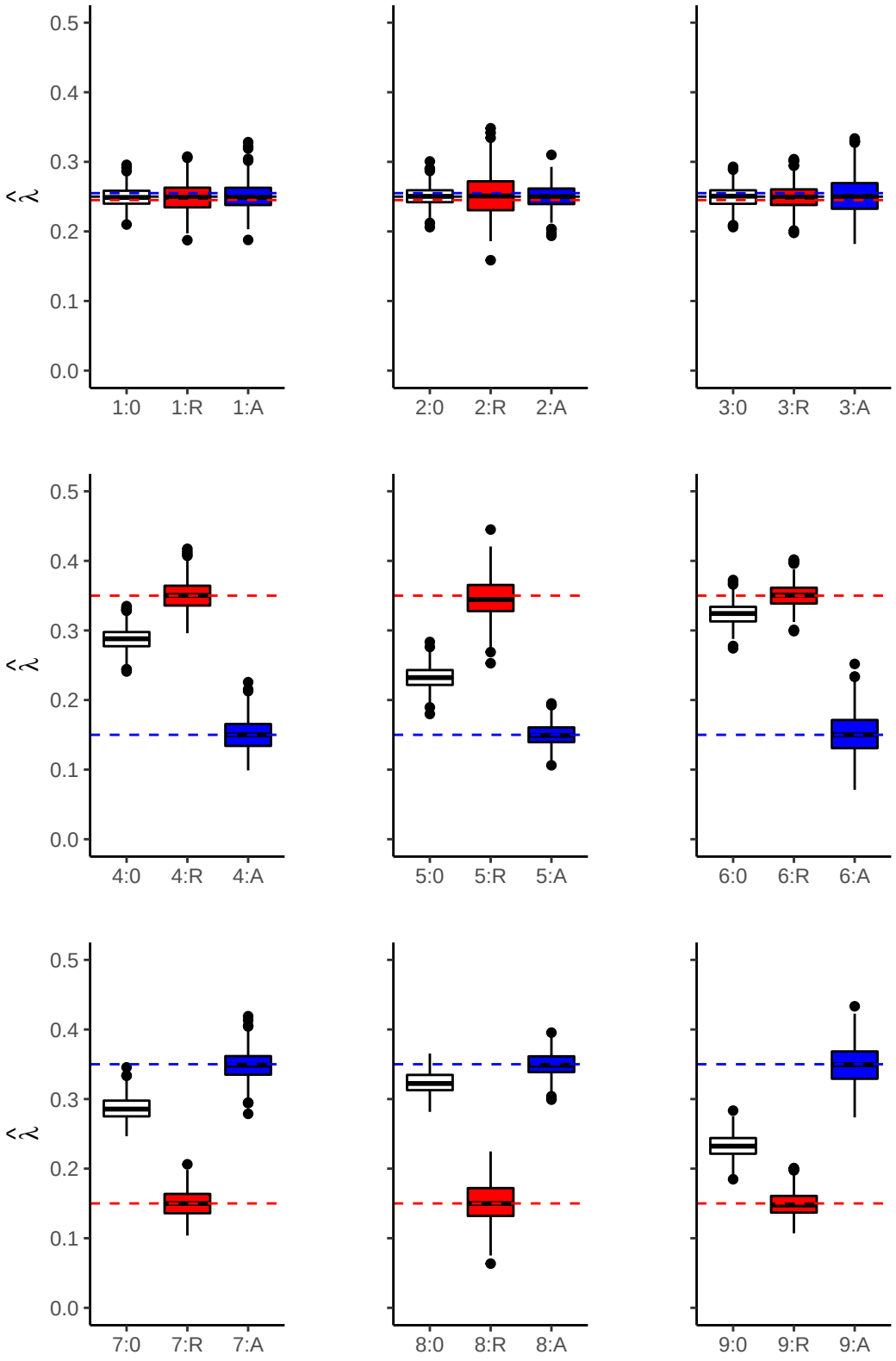

Figure 3. Boxplots displaying the distribution of the Poisson parameter estimates for simulation 1.

Note. Dashed lines depict true Poisson parameters.

3.2 Simulation 2

In the second simulation, the police district in which the capture took place was simulated as a time-varying event-related covariate. The district-specific rates and proportion of time that was spent in each district were rotated in nine conditions, presented in Table 3.

Table 3

Conditions in Simulation 2

Condition	N	r_R	r_A	t_R	t_A	λ_R	λ_A	λ_{R+A}
1	10,000	0.0025	0.0025	50	50	0.1250	0.1250	0.250
2	10,000	0.0030	0.0020	50	50	0.1500	0.1000	0.250
3	10,000	0.0020	0.0030	50	50	0.1000	0.1500	0.250
4	10,000	0.0025	0.0025	75	25	0.1875	0.0625	0.250
5	10,000	0.0030	0.0020	75	25	0.2250	0.0500	0.275
6	10,000	0.0020	0.0030	75	25	0.1500	0.0750	0.225
7	10,000	0.0025	0.0025	25	75	0.0625	0.1875	0.250
8	10,000	0.0030	0.0020	25	75	0.0750	0.1500	0.225
9	10,000	0.0020	0.0030	25	75	0.0500	0.2250	0.275

Note. λ_{R+A} is the summed Poisson parameter of observation period $T = 100$. λ_R and λ_A represent the respective Poisson parameters of subperiods t_R and t_A.

3.2.1 Point and interval estimates of N. The boxplots in Figure 4 show the performance of the four methods through the distribution of the population size estimates in the nine conditions of the second simulation. Table 4 reports the corresponding coverage probabilities .

The intercept-only model accurately estimated the total population size in all nine conditions, which is depicted in the first nine rows of Figure 4. While deviating from its true size in some cases, the average of the estimates was close to the true population size.

This accurate estimation was further confirmed by the coverage probabilities, which ranged from 93% to 96%, but evened out around 95% overall.

The second nine rows of Figure 4 show the estimates of the one dummy model. While the estimates of the other models were somewhat stable, those of the one dummy model varied. Though the average of the population size estimates was too high, there were some cases in which its estimates were close to their true size. Moreover, the coverage probabilities of this model varied substantially, ranging from 32% in condition 7 to 93% in condition 5. Nevertheless, the coverage probabilities did not approach the nominal coverage probability of 95% on the whole.

The performance of the two dummy and count models is displayed in the lower half of Figure 4. Again, these models overestimated the total population size in all conditions. Consequently, their coverage probabilities were close to zero in all nine conditions, which is subnominal.

Table 4

95% Coverage Probabilities per Condition for All Methods in Simulation 2

	Coverage probabilities			
Condition	Intercept-only	One dummy	Two dummies	Count
1	0.950	0.548	0.000	0.000
2	0.944	0.724	0.000	0.000
3	0.934	0.416	0.000	0.000
4	0.944	0.902	0.000	0.000
5	0.960	0.926	0.032	0.000
6	0.956	0.820	0.000	0.000
7	0.948	0.318	0.000	0.000
8	0.948	0.504	0.000	0.000
9	0.960	0.370	0.022	0.000

3.2.2 Point estimates of λ. The boxplots in Figure 5 show the estimates of the Poisson parameter for the intercept-only (0) and one dummy models (R and A). Similar to the first simulation, the line marks the true value of the Poisson parameter, the boxplots represent what was estimated.

As can be gathered from the figures, the estimates of the intercept-only model were unbiased in all nine conditions. Moreover, this model produced stable estimates of the Poisson parameter with little variation in the spread of the values across the conditions.

The results are less readily described for the one dummy model, which strayed from the true Poisson parameter in all conditions. Remarkably, the estimates of the constant deviated most when the proportion of time that was spent in district R was smallest ($t_R = 25$). Likewise, when the proportion of time that was spent in district A was smallest ($t_A = 25$), the estimates of the dummy strayed most from the true Poisson parameter. Lastly, compared to the intercept-only model, the one dummy model showed more variation in its estimates of the Poisson parameters as the boxplots are slightly bigger, indicating that the values are spread further apart.

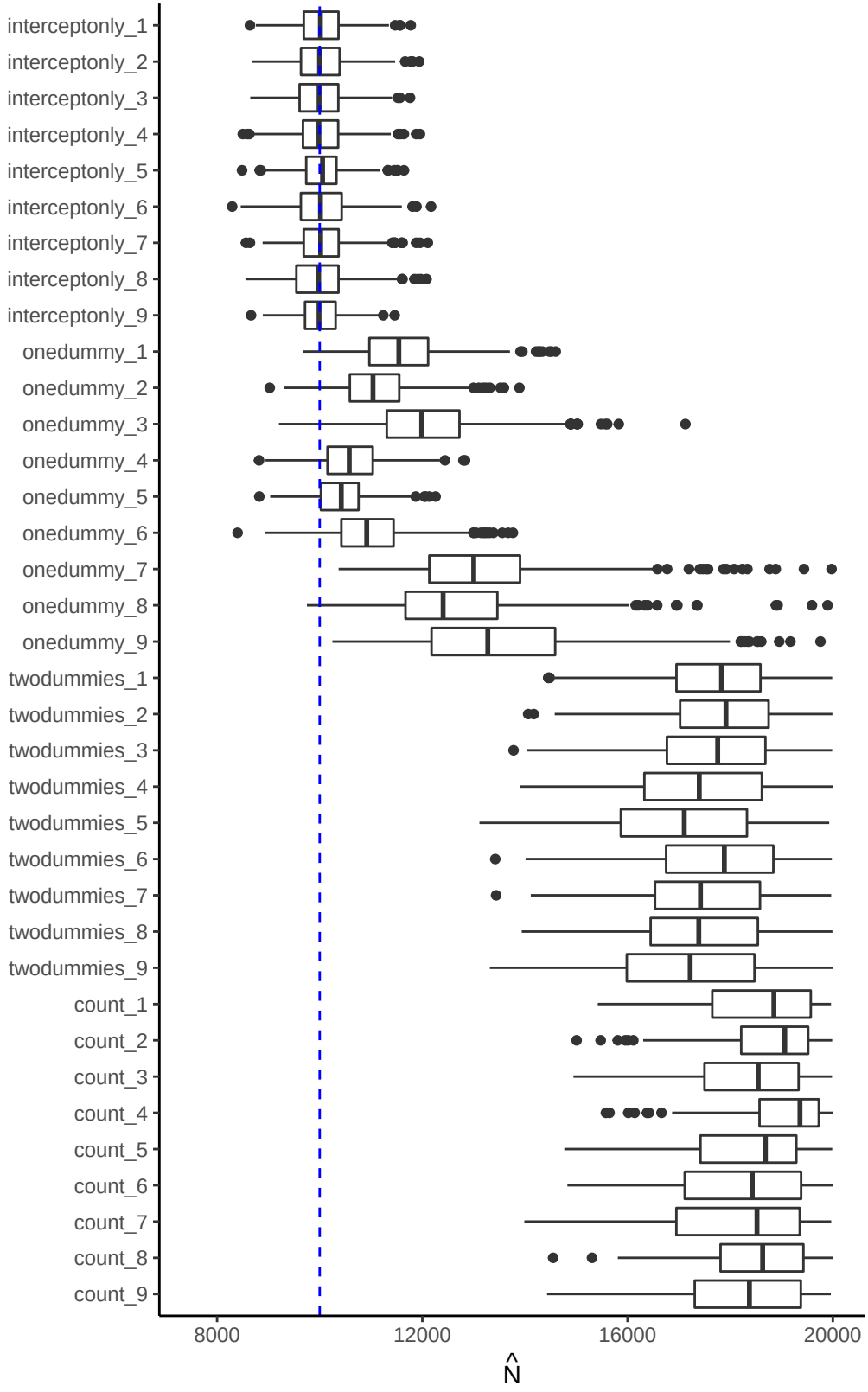

Figure 4. Boxplots displaying the distribution of the population size estimates for simulation 2.

Note. Dashed line depicts true total population size.

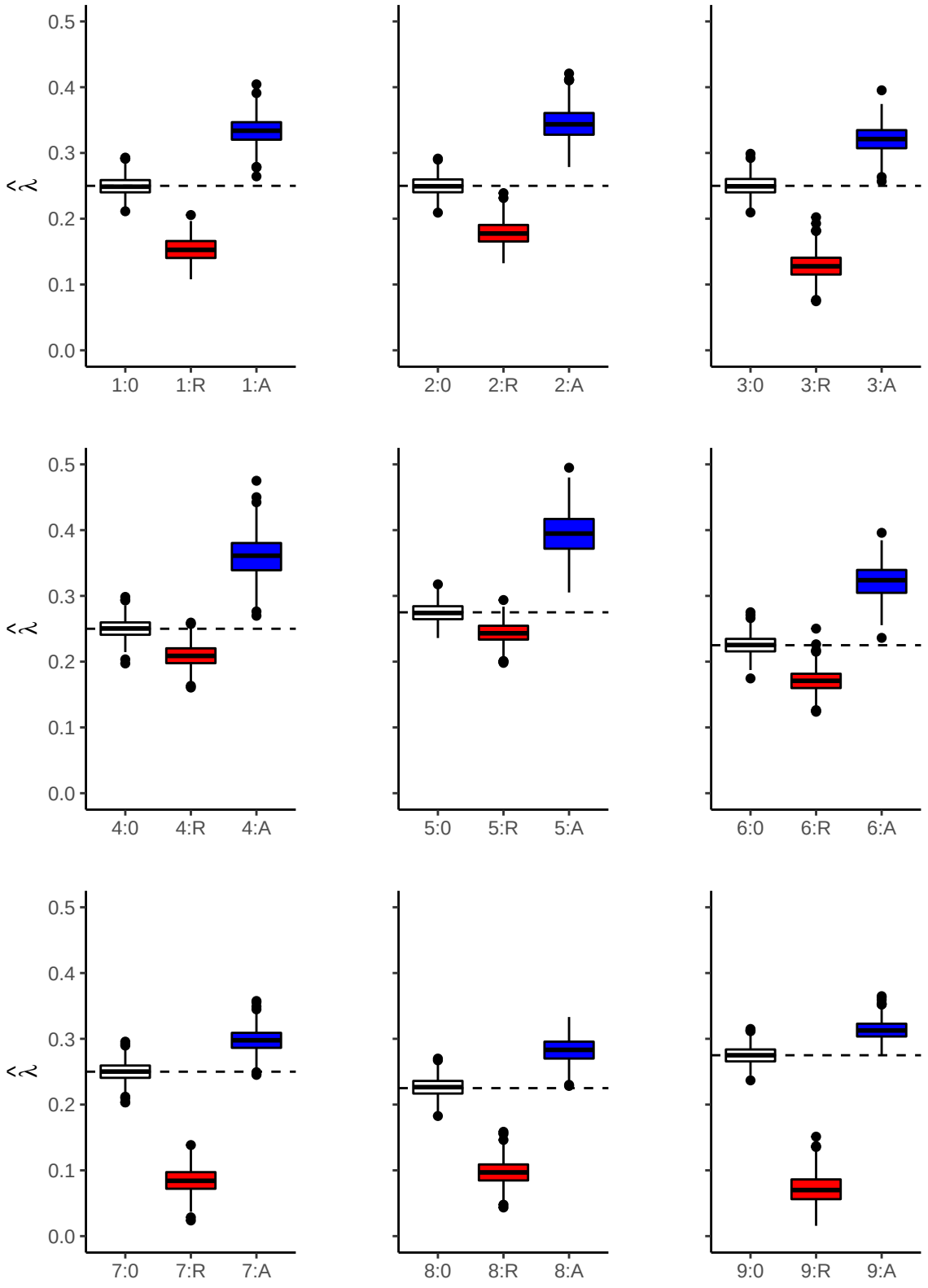

Figure 5. Boxplots displaying the distribution of the Poisson parameter estimates for simulation 2.

Note. Dashed line depicts true Poisson parameter.

3.3 Simulation 3

The first and second simulations were combined in the third simulation in which the event-related covariate was considered to be both invariant and time-varying. Table 5 shows the conditions that were considered. Since the two dummy and count models consistently overestimated the total population size in the previous simulations, these models were omitted from the third simulation.

Table 5

Conditions in Simulation 3

	Condition	r_R	r_A	t_R	t_A	T	λ_R	λ_A	λ_{A+R}
Subpopulation 1	1	0.0025	0.0025	-	-	100	0.25	0.25	-
	2	0.0030	0.0020	-	-	100	0.30	0.20	-
	3	0.0020	0.0030	-	-	100	0.20	0.30	-
	4	0.0025	0.0025	-	-	100	0.25	0.25	-
	5	0.0030	0.0020	-	-	100	0.30	0.20	-
	6	0.0020	0.0030	-	-	100	0.20	0.30	-
	7	0.0025	0.0025	-	-	100	0.25	0.25	-
	8	0.0030	0.0020	-	-	100	0.30	0.20	-
	9	0.0020	0.0030	-	-	100	0.20	0.30	-
Subpopulation 2	1	0.0025	0.0025	50	50	100	0.125	0.125	0.250
	2	0.0030	0.0020	50	50	100	0.150	0.100	0.250
	3	0.0020	0.0030	50	50	100	0.100	0.150	0.250
	4	0.0025	0.0025	75	25	100	0.1875	0.0625	0.250
	5	0.0030	0.0020	75	25	100	0.225	0.050	0.275
	6	0.0020	0.0030	75	25	100	0.150	0.075	0.225
	7	0.0025	0.0025	25	75	100	0.0625	0.1875	0.250
	8	0.0030	0.0020	25	75	100	0.075	0.150	0.225
	9	0.0020	0.0030	25	75	100	0.050	0.225	0.275

Note. Subpopulation 1 consists of two groups with $n_{1,2} = 2,500$, and Subpopulation 2 of one group with $n_3 = 5,000$; $N = 10,000$. For Subpopulation 2, λ_{R+A} is the summed Poisson parameter of observation period T. λ_R and λ_A represent the respective Poisson parameters of subperiods t_R and t_A.

3.3.1 Point and interval estimates of N. The boxplots in Figure 6 show the performance of the two models through the distribution of the population size estimates in the nine conditions of the third simulation. Table 6 reports the corresponding coverage probabilities.

As can be gathered from Figure 6, the intercept-only model produced accurate estimates of the total population size when the rates were homogeneous in conditions 1, 4, and 7. In all other conditions, both the intercept-only and one dummy models over- or underestimated the size of the total population. Despite this, the coverage probabilities of both models were close to the nominal level of 95%. When the models are compared, two things about the intercept-only model stand out. First, its estimates were closer to the true population size. Second, its estimates were more stable.

Table 6

95% Coverage Probabilities per Condition for All Methods in Simulation 3

	Coverage probabilities	
Condition	Intercept-only	One dummy
1	0.954	0.942
2	0.916	0.926
3	0.912	0.882
4	0.942	0.952
5	0.924	0.928
6	0.894	0.934
7	0.932	0.928
8	0.918	0.928
9	0.924	0.904

3.3.2 Point estimates of λ. The boxplots in Figure 7 show the estimates of the Poisson parameters for the intercept-only (0) and one dummy model (R and A). Again,

the colored lines mark the true values of the Poisson parameters, the boxplots represent what was estimated.

The estimates of the Poisson parameter of the intercept-only model were unbiased when the rates were homogeneous in conditions 1, 4, and 7. In all other conditions, the estimates slightly deviated from the true Poisson parameters. Similar to the previous simulations, the estimates were stable, with relatively little variation in their spread. In contrast, the one dummy model produced biased estimates of the Poisson parameters in all nine conditions. For homogeneous rates, the estimates were disparate, resulting in an overestimation of the total population size. Comparatively, for heterogeneous rates, when the rate of district R was higher than that of district A, the estimates were comparable. This coincided with an underestimation of the total population size. In contrast, when the rate of district A was higher than that of district R, the estimates varied, which corresponded to an overestimate of the total population size by a larger margin than for homogeneous rates.

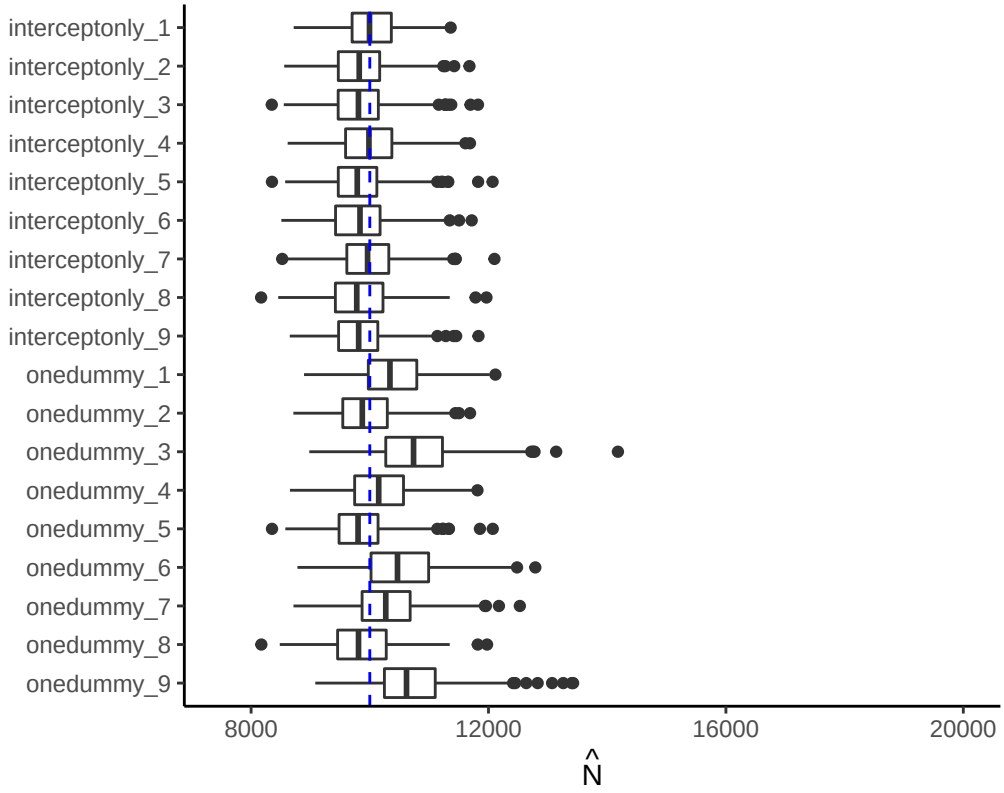

Figure 6. Boxplots displaying the distribution of the population size estimates for simulation 3.

Note. Dashed line depicts true total population size.

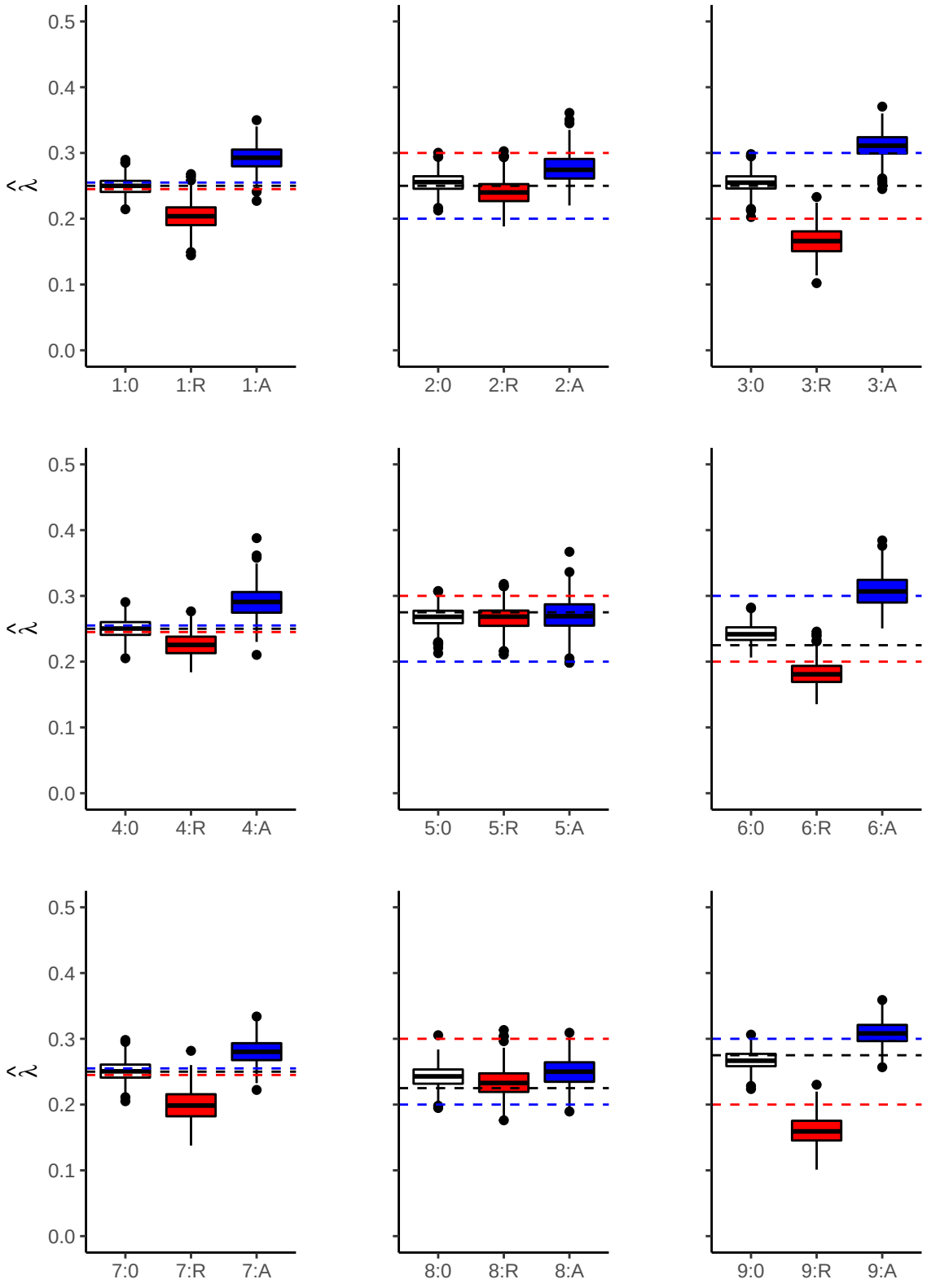

Figure 7. Boxplots displaying the distribution of the Poisson parameter estimates for simulation 3.

Note. Dashed lines depict true Poisson parameters.

4 Conclusion and Discussion

This study evaluated the performance of multiple methods for including an event-related covariate in the zero-truncated Poisson regression model in three simulations. In the first and second simulation, the event-related covariate was respectively considered invariant and time-varying. This distinction was discarded in the third simulation, where the event-related covariate manifested itself as both invariant and time-varying.

The count and two dummy models performed unsatisfactorily as the total population size was overestimated for both invariant and time-varying event-related covariates. Next, the results of the one dummy and intercept-only models varied based on whether the value of the event-related covariate could change over time. When its value was invariant, the one dummy model accurately estimated the Poisson parameters and population size. In contrast, the intercept-only model only produced accurate estimates when the rates were homogeneous. Further, when the event-related covariate was regarded as time-varying, the estimates of the one dummy model were biased, whereas those of the intercept-only model were accurate. Lastly, both models performed comparable in the mixture model, with the population size estimates of the intercept-only model being closer to the true size and less fluctuate.

Taken together, the results of this simulation study with these conditions are twofold. On the one hand, when an event-related covariate is regarded as invariant, it should be included with the use of a dummy. On the other hand, when an event-related covariate is at least partially considered to vary over time, it should not be included in the model at all. Their inclusion in any way ensues biased estimates of the Poisson parameters and the

total population size.

To clarify these results, we retract to the properties of the zero-truncated Poisson regression model. As described in Section 2.2, covariates divide the total population into subgroups. For an estimate of a population to benefit from the inclusion of covariates, these subgroups should be meaningful, i.e. the covariates should contribute significantly to the proportion of observed heterogeneity. This is the case when the value of the event-related covariate is invariant, because different subgroups can be distinguished based on the individual characteristic it then represents. However, when the value of the event-related covariate at least partially varies over time with the population members divided over the categories in a similar manner, there are no subgroups to differentiate between. The heterogeneity then lies outside the population, as the event-related covariate does not represent an individual characteristic, but a property of the event count under study. Consequently, the heterogeneous Poisson parameters may be summed to form one homogenous Poisson parameter, as described in Section 2.3.

When the event-related covariate is included in the model regardless, this results in a special case of Jensen's inequality as described in Section 2.2, where the covariate specifies subgroups that ensure a higher expectation of a zero count. However, as the subgroups are spurious, the higher expectation of a zero count is in fact an overestimation that emanates the subsequent overestimation of the total population size. Then, as the two dummy and count models distinguish between even more spurious subgroups, these models overestimate the total population size by a larger margin than the one dummy model.

Although it is difficult to aggregate these findings to population size estimation in

general as the results of these simulations are contingent on their conditions, an assumption should be made about the variation over time of the event-related covariate. Based on that assumption, the method for including event-related covariates should be determined. Within the present study, the mixed population of the third simulation describes the most realistic situation. Considering the estimates of the intercept-only model in this simulation were more stable, that is to say, more predictable, we would discourage the approach taken by Van Der Heijden et al. (2003b) to include event-related covariates with dummies. Instead, event-related covariates should not be taken into account as covariates in the zero-truncated Poisson regression model. The population size estimate could then function as a lower bound.

However, there are some limitations to the current research. First of all, the simulation consisted of a limited number of conditions, ergo the generalizability of this study is restricted. Moreover, the event-related covariate in the simulations had only two categories. It would be interesting to investigate how the results hold up in more complex situations, like when the number of values and categories is expanded upon. Additionally, the effect of different group movements across categories was not studied, as this laid outside the scope of the present paper. Instead, when the event-related covariate was considered time-varying, the population members were assumed to divide their time between the police districts in a similar manner. Hereby, potential subgroups could not be distinguished based on time allocation. Future research should diversify the variation over time within the population and see whether the results are comparable. Besides, the present study did not consider violations of the assumptions of the Poisson model. For

example, we assumed that every population member had a Poisson parameter, and that within that Poisson parameter, rate and capture were independent. What is more, we presumed that the event-related covariate was the only source of heterogeneity within the population. Consequently, there was no unobserved heterogeneity that could cause an underestimation of the total population size. Further research should try to approximate these violations of the model in a simulation study and examine how this affects the results.

Despite these limitations, the present paper has contributed to the existing body of literature on the methods for population size estimation in providing evidence that one should be careful in taking event-related covariates into account as a source of heterogeneity. Worse yet, when these covariates are wrongfully included, they are likely to bias the estimates. While further research should be conducted to solidify this finding, the current study has raised awareness of the complexity of including event-related covariates in the zero-truncated Poisson regression model, and its implications should be contemplated in attempts to estimate the size of elusive populations.

References

Aust, F., & Barth, M. (2020). *papaja: Create APA manuscripts with R Markdown*. Retrieved from https://github.com/crsh/papaja

Austin, P. C., Latouche, A., & Fine, J. P. (2020). A review of the use of time-varying covariates in the fine-gray subdistribution hazard competing risk regression model. *Statistics in Medicine*, *39*(2), 103–113. https://doi.org/10.1002/sim.8399

Bonner, S. J., Morgan, B. J., & King, R. (2010). Continuous covariates in mark-recapture-recovery analysis: A comparison of methods. *Biometrics*, *66*(4), 1256–1265. https://doi.org/10.1111/j.1541-0420.2010.01390.x

Bonner, S. J., Thomson, D. L., & Schwarz, C. J. (2009). Time-varying covariates and semi-parametric regression in capture–recapture: An adaptive spline approach. In *Modeling demographic processes in marked populations* (pp. 657–675). Springer. https://doi.org/10.1007/978-0-387-78151-8_29

Cameron, A. C., & Trivedi, P. K. (1998). *Regression analysis of count data* (Vol. 53). Cambridge, United Kingdom: Cambridge University Press.

Catchpole, E. A., Morgan, B. J., & Tavecchia, G. (2008). A new method for analysing discrete life history data with missing covariate values. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *70*(2), 445–460. https://doi.org/10.1111/j.1467-9868.2007.00644.x

Cortese, G., & Andersen, P. K. (2010). Competing risks and time-dependent covariates. *Biometrical Journal*, *52*(1), 138–158. https://doi.org/10.1002/bimj.200900076

Cruyff, M. J., & Van Der Heijden, P. G. (2008). Point and interval estimation of the

population size using a zero-truncated negative binomial regression model. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, *50*(6), 1035–1050. https://doi.org/10.1002/bimj.200810455

Hinde, J., & Demétrio, C. G. (1998). Overdispersion: Models and estimation. *Computational Statistics and Data Analysis*, *27*(2), 151–170. https://doi.org/10.1016/S0167-9473(98)00007-3

Hoogteijling, E. (2002). *Raming van het aantal niet in de gba geregistreerden*. The Hague, The Netherlands: Statistics Netherlands. Retrieved from https://www.cbs.nl/nl-nl/achtergrond/2002/16/raming-van-het-aantal-niet-in-de-gba-geregistreerden

Johnson, N. L., Kemp, A. W., & Kotz, S. (2005). *Univariate discrete distributions* (Vol. 444). New York, United States: John Wiley & Sons.

Kalbfleisch, J., & Prentice, R. L. (1980). *The statistical analysis of failure time data*. New York, United States: Wiley.

Kendall, W. L., Conn, P. B., & Hines, J. E. (2006). Combining multistate capture–recapture data with tag recoveries to estimate demographic parameters. *Ecology*, *87*(1), 169–177. https://doi.org/10.1890/05-0637

Kingman, J. F. C. (1993). *Poisson processes*. New York, United States: Oxford University Press.

Kromhout, M., Wubs, H., & Beenakkers, E. (2008). *Illegaal verblijf in nederland*. The Hague, The Netherlands: Wetenschappelijk Onderzoek- en Documentatiecentrum. Retrieved from https://www.wodc.nl/binaries/cahier-2008-3-volledige-tekst_tcm28-70022.pdf

Lancaster, T. (1990). *The econometric analysis of transition data*. Cambridge, United Kingdom: Cambridge University Press.

Lovett, A., & Flowerdew, R. (1989). Analysis of count data using poisson regression. *The Professional Geographer*, *41*(2), 190–198. https://doi.org/10.1111/j.0033-0124.1989.00190.x

Meschiari, S. (2015). *Latex2exp: Use latex expressions in plots*. Retrieved from https://CRAN.R-project.org/package=latex2exp

Nelder, J. A., & Wedderburn, R. W. (1972). Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, *135*(3), 370–384. https://doi.org/10.2307/2344614

Raikov, D. (1938). On the decomposition of gauss and poisson laws (in russian). *Izvestiya Rossiiskoi Akademii Nauk. Seriya Matematicheskaya*, *2*(1), 91–124. Retrieved from http://www.mathnet.ru/php/archive.phtml?jrnid=im&wshow=contents&option_lang=eng

R Core Team. (2020). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Rostgaard, K. (2008). Methods for stratification of person-time and events–a prerequisite for poisson regression and sir estimation. *Epidemiologic Perspectives & Innovations*, *5*(1), 7. https://doi.org/10.1186/1742-5573-5-7

Snippe, J., & Mennes, R. (2018). Vooronderzoek data en methoden illegalenschatting. Groningen, The Netherlands: Breuer en Intraval Onderzoek & Advies. Retrieved

from https://www.wodc.nl/binaries/2917_Volledige_Tekst_tcm28-356573.pdf

"Student". (1919). An explanation of deviations from poisson's law in practice. *Biometrika*, *12*, 211–215. https://doi.org/10.2307/2331767

Upton, G., & Cook, I. (1996). *Understanding statistics*. Oxford, United Kingdom: Oxford University Press.

Van Der Heijden, P. G., Bustami, R., Cruyff, M. J., Engbersen, G., & Van Houwelingen, H. C. (2003a). Point and interval estimation of the population size using the truncated poisson regression model. *Statistical Modelling*, *3*(4), 305–322. https://doi.org/10.1191/1471082x03st057oa

Van Der Heijden, P. G., Cruyff, M., & Van Houwelingen, H. C. (2003b). Estimating the size of a criminal population from police records using the truncated poisson regression model. *Statistica Neerlandica*, *57*(3), 289–304. https://doi.org/10.1111/1467-9574.00232

Vermunt, J. K. (1996). *Log-linear event history analysis: A general approach with missing data, latent variables, and unobserved heterogeneity* (Vol. 8). Tilburg, The Netherlands: Tilburg University Press.

Wickham, H. (2016). *Ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York. Retrieved from https://ggplot2.tidyverse.org

Wilke, C. O. (2019). *Cowplot: Streamlined plot theme and plot annotations for 'ggplot2'*. Retrieved from https://CRAN.R-project.org/package=cowplot

Yamaguchi, K. (1986). Alternative approaches to unobserved heterogeneity in the analysis of repeatable events. *Sociological Methodology*, *16*, 213–249.

https://doi.org/10.2307/270924