Modern Regression Methods, MATH-408

Analysis of litter on Swiss shores.

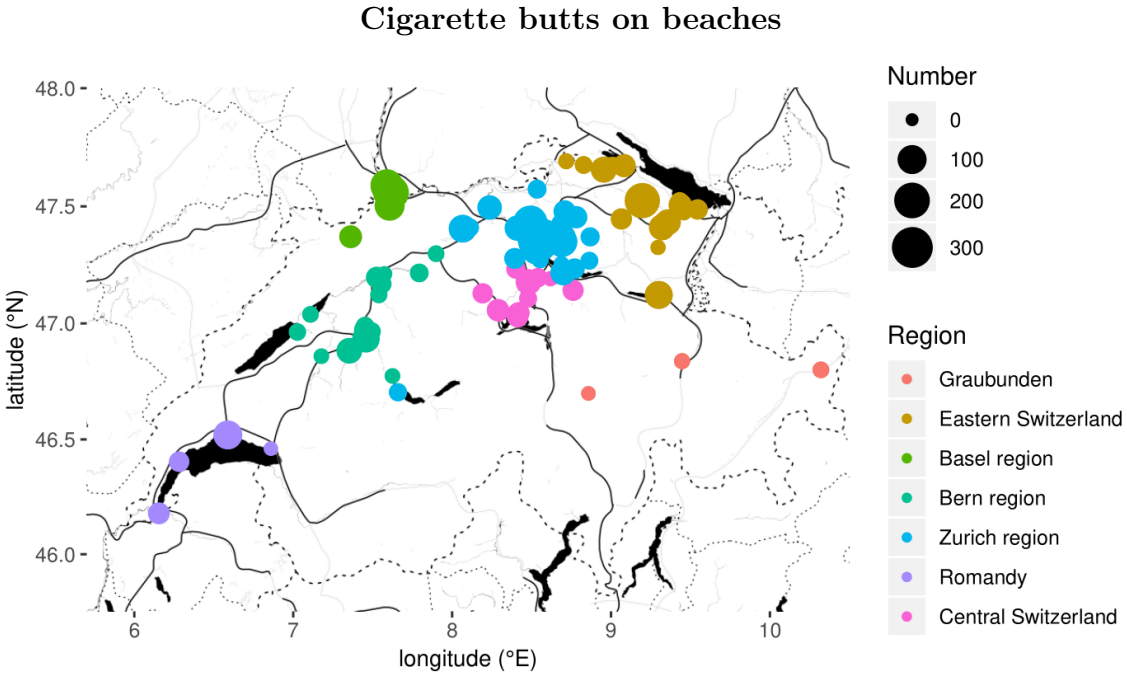
—

William Cappelletti, Charles Dufour

Abstract

We analyze a data set on cigarette butts count on different shores in Switzerland. We fit three different Generalized Linear Models and a Generalized Additive Model with Negative Binomial distribution using various geographical and temporal information. Then, we extensively discuss our final model, observing an interesting seasonal effect and some regional peculiarities. Finally, we use our results to give confidence intervals for the cigarette butts count in the month of August 2019 for all the considered beaches, with a particular attention to those on the shores of Geneva Lake. We regress their expected mean by smoothing on both the months and the inverse of the length of the beach, accounting for a fixed effect given by the Region and a random effect from the beach.

Keywords: Litter, Cigarettes, Pollution, Lakes, Rivers, Urbanization



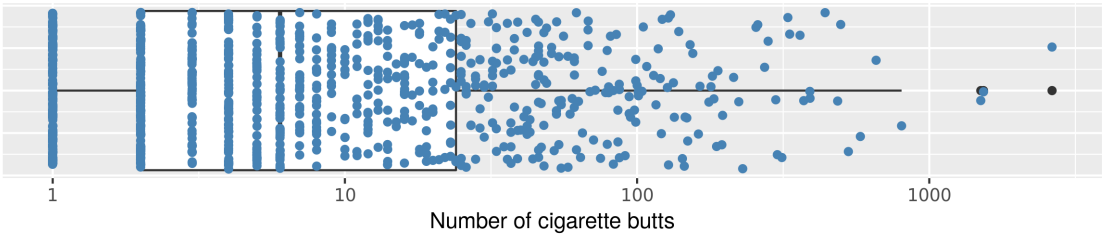


Figure 1: Cigarette butts number Boxplot. We can see a few outliers.

1 Introduction

In this report we study the number of cigarette butts collected by trained volunteers on the shores of various Swiss lakes and shores. These data are provided by [2]. We aim to find an insightful model able to give a prediction interval for the number of butts, our response, on a given beach in a given month, say August 2019. To do so, we analyse 855 observations of 109 beaches across seven Swiss Regions over 12 month; from April 2017 to March 2018. These data are the synthesis of 905 observations, in which some shores were cleaned more than once per month; to get the monthly count we thus sum the observations of the same month. Furthermore, for each beach we know its length, whether it is in a *city*, in an *agglomeration* or in the *country* and, for many of them, an estimated number of people who go there and the average fraction of the shore washed by the waves.

In Section 2, we explore the data-set, in order to understand the nature of the response and to assess the presence of any outlier. Figure 1 shows a box-plot of the cigarette butts count, in which the response seems to follow some sort of Poisson distribution, which is in line with the fact that we are working with count values. Consequently, in Section 3 we fit different models, starting with Generalized Linear Models (GLM) with a Poisson distribution, to move to Generalized Additive Models (GAM), in which we focus on quasi-Poisson and Negative-Binomial distribution to account for over-dispersion. Then, in Section 4, we discuss the obtained results and the performance of our model in the predictive task.

2 Preliminary data analysis

A quick overview on the data-set points out a beach with an absurd length of 1335049 meters, thus we discard the two observations linked to it, leaving us with 855 data-points and 109 different beaches. One of this beaches, namely one on river Sihl close to Leimbach is reported with two different lengths (27 and 99 meters); by checking satellite images, we see that the correct length is 99 meters, thus we manually change the measure.

Then, the boxplot in Figure 1 shows that only few observations exceed 1000 counts, while most of the observations are close to zero. These extreme values come from the same beach, namely the "Rhein Beach" near the "Tinguely Museum" in Basel, which is in a highly urbanized area, close to some big streets. The nature of this shore could explain such a high number of litter, but we discard it as an outlier since it is the only one showing this extreme behaviour.

Table 1 contains the quartiles of the number of cigarette butts and their frequency, defined as the butts count over the length of the beach. Based on that we choose to drop as outliers all the observations with a frequency greater than four, thus losing 28 observations.

As we previously said, Figure 1 shows that the data are almost Poisson distributed. Figure 3,

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Number	0.00	1.00	5	26.07	22.00	657
Frequency	0.00	0.04	0.18	0.68	0.72	13.14

Table 1: Quantiles of Cigarette butts number and frequency.

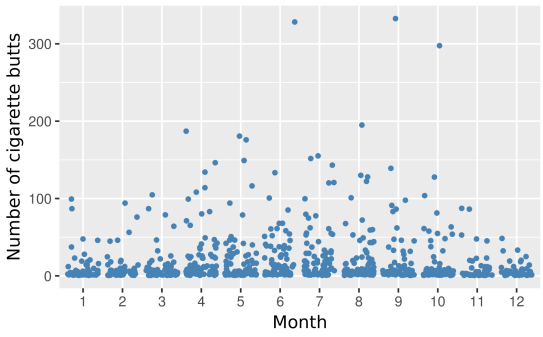


Figure 2: Number of cigarette butts over different months.

Figure 3: Number of cigarette butts divided by Region of the beach.

which shows the number of cigarette butts divided by the Region of the beach, supports this claim and shows that the region influence the variance of the response.

Table 2 reports the correlations between the main covariates and the response. We can see that some of them seem very correlated, which is a consequence of the way the observations were gathered and their structure. In fact, all the geographical information, i.e. the beach name, the region, and the type, are uniquely paired. Furthermore, the months do not overlap over different years, thus explaining the correlation over the time measures; this could suggest removing from our analysis the yearly component, as a difference in the seasons could be interpreted by the model as a difference between 2017 and 2018. The number of people going to the beach and the shore's surface washed by the waves are ignored since they have a very low correlation with the response and they have many missing values, which require to drop other observations.

	Beach	Length	Year	Month	Region	Type	Number
Beach	*	–	–	–	–	–	–
Length	1.00	*	–	–	–	–	–
Year	0.25	0.03	*	–	–	–	–
Month	0.22	0.08	1.00	*	–	–	–
Region	1.00	0.23	0.05	0.07	*	–	–
Type	1.00	0.05	0.03	0.07	0.50	*	–
Number	0.73	0.14	0.12	0.20	0.27	0.22	*

Table 2: Correlation table between the covariates and the response.

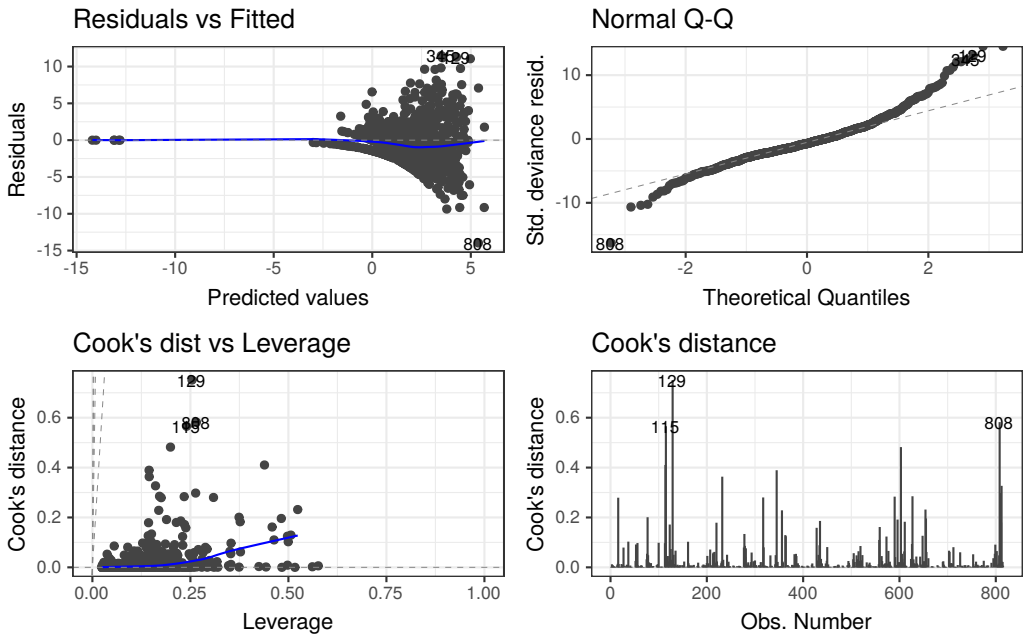


Figure 4: Diagnostics for the Poisson GLM.

3 Analysis and Modelling

Since we are dealing with count data, we start by fitting a Generalized Linear Model (GLM), using the R function `glm`, with a Poisson distribution and its canonical link function log. More precisely, we suppose that the number of butts y is distributed as

$$y \sim \text{Poiss}(l\mu), \qquad \log(\mu) = x^{\mathrm{T}}\beta, \tag{1}$$

where l is the length of the beach, here used as offset, x is the covariate's vector, and β are the linear coefficients. We choose to use the beach length as offset as it seems logical that this covariate only increases the effect of the others. In fact, we imagine that, given all other factors, a scaling of the beach would scale accordingly the response.

Since the beaches are more frequented during summer than winter we suppose that there is a seasonal effect over the month, which is supported by Figure 2. It seems that the months modify the variance of the response, in particular it increases during summer and decreases in winter. Because of that we fit a seasonal component over the month m, by adding as covariates the first two terms of the Fourier series, namely

$$\sin\left(\frac{2\pi m}{12}\right),\ \cos\left(\frac{2\pi m}{12}\right),\ \sin\left(\frac{4\pi m}{12}\right),\ \cos\left(\frac{4\pi m}{12}\right).$$

Our first model is based on (1) and it uses all the previously cited parameters. We fit it in R by calling

```
glm(formula = Number ~ sin(2 * pi * m/12) + cos(2 * pi * m/12) +
      sin(4 * pi * m/12) + cos(4 * pi * m/12) + Region + Type + Beach +
      Year + offset(log(Length)) - 1, family = poisson, data = df)
```

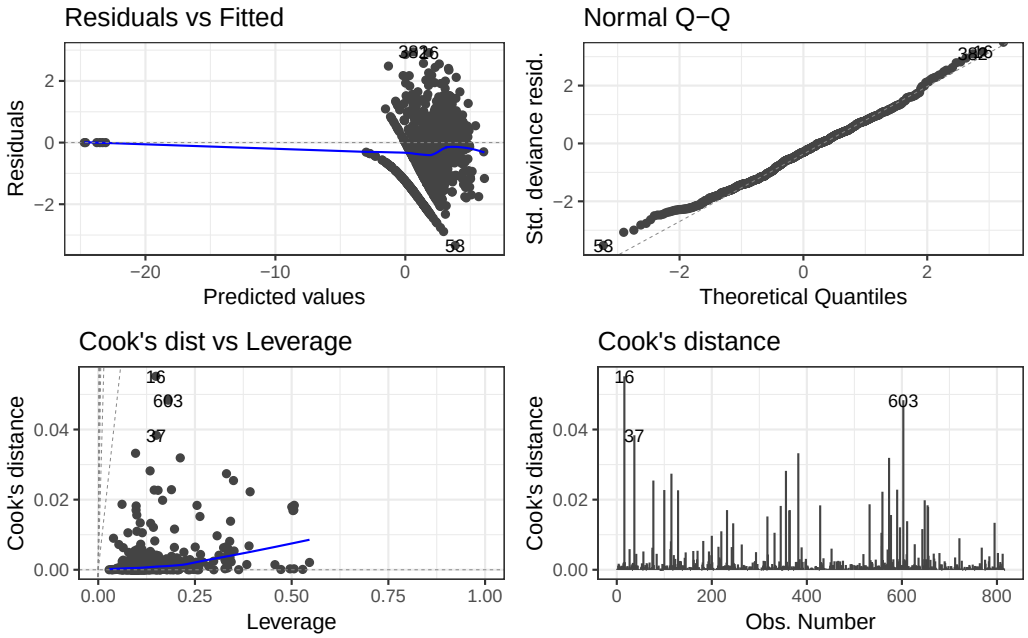


Figure 5: Diagnostics for the Negative Binomial GLM.

This model has a deviance of 6877.7 and AIC of 9779.1. By backward selection based on AIC, we see that a model without the covariate 'type' gives the same deviance and the same information criteria. Therefore we stick to the more parsimonious model and Figure 4 shows its diagnostics. On the top left we see that the residuals against the predicted values show signs of overdispersion, which is supported by the QQ-plot on the top right, where the tails of the standardised deviance residuals are heavier than the theoretical ones. The two lower sub-Figures show the Cook's distance against the leverage and the observation number. We see three observations with a high statistic, but they have an average leverage and, by closely looking at the data, they do not seem to be symptoms of a bad fit.

In light of these observations, we move to fit some models which can account for overdispersion. In particular we focus on quasi-Poisson and Negative Binomial. The quasi-Poisson model arise from the following second order assumptions, in addition to the hypothesis of (1)

$$\mathbb{E}[Y] = l\mu, \ \operatorname{var}(Y) = \phi \operatorname{var}(\mu), \ \log(\mu) = x^{\mathrm{T}}\beta. \tag{2}$$

By doing the same experiment as before and specifying the quasi-Poisson family in the `glm` formula, we obtain the same deviance as before, as well as almost indistinguishable diagnostics. We therefore move to a different model.

'+ The Negative Binomial model arise by supposing that, conditional on a gamma distributed variable ϵ,

$$Y \sim \text{Poiss}(l\mu\epsilon), \ \log(\mu) = x^{\mathrm{T}}\beta, \tag{3}$$

where l is again the beach legth used as offset. Again, we fit a first model using all the explanatory variables, namely the Fourier basis of the month, the region, the type, the beach, and the year. This model has an AIC of 5142.2 and a deviance of 900.3. From that, we perform a backward selection based on AIC reduction and we drop the covariates 'region', 'type', and $sin(2\pi m/12)$. We end up with a model giving an AIC of 5141 and a deviance of 901.7. Figure 5 shows some diagnostics of the step wise reduced model. The residuals in the top-left sub-Figure are less spread than before, all

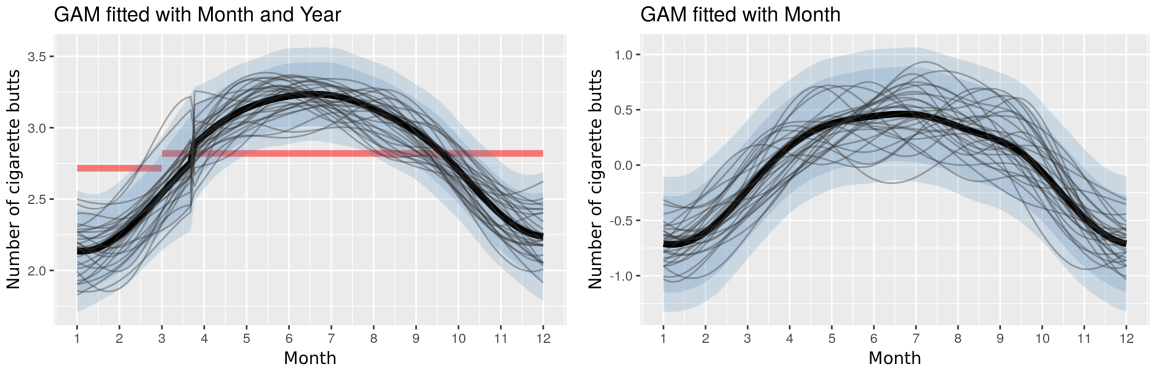


Figure 6: Point-wise and Simultaneous 95% confidence intervals for a GAM fitted by smoothing the months using cyclic cubic splines. Each blue line is one of 30 draws from the Bayesian posterior distribution of the model and the thick black line is the mean.
Left panel is fitted including the year, which estimated effect is shown in red, while right panel only uses the month smoothing.

being in the interval $[-3, 3]$. In addition, the Q-Q plot in the top right panel is quite well behaved, but it still shows a heavier tail in the left side. The two Cook's distance plots only point out three data point with a high statistic; they are not the same as before, but nonetheless their leverage is not so high and there does not seem to be any oddities in their figures.

The Negative Binomial model we just obtained is the better so far, thus suggesting a good choice of distribution. Nevertheless, we try to add on top of that a further thought on the relations between the covariates. In particular, we wonder whether the introduction of random effects and smoothing could improve our model. Therefore, we move from GLMs to Generalized Additive Models (GAMs), which we fit using the R function `gam` with the REML optimization routine. Our first idea is to consider the beaches themselves as a random effect, since they do not interest us by themselves, but we would like to consider them as representatives of a greater population. Furthermore, we rethink the use of the covariates 'year' and 'month'.

As we already said in Section 2, in 2017 we only observe from April to December, while in 2018 our observations are gathered from January to March. This could imply that a part of the variation in the first three months of the year will be absorbed by the year variable, while it could be more logical to consider it as a seasonal effect. The two panels of Figure 6 show that it is not the case. They both show the smoothing we obtain by fitting cyclic cubic splines, thus forcing seasonality, over the month in a model with, and respectively without, the 'year' as an explanatory variable. As we can see, both models produce confidence intervals (CI) with almost constant width all over the year, but the inclusion of the yearly component slightly shrinks the CI. We thus choose to include from the beginning the year as a random effect, as we have no evidence of the presence of a fixed effect or some kind of trend, and we would like to be able to generalize the model to unobserved years.

The choice of smoothing on the month instead of using a sinus and cosinus expansion is supported by the AIC, which is slightly lower in the former model when the other covariates are the same. For instance, in the full model, we have that the AIC is 5144.1 when using the explicit expansion, while it is 5142.5 when smoothing. Therefore, we use the smoothed model, as it has a nicer interpretation. We also see that it is not useful to separately smooth the months over the regions, as it highly increases the complexity of the model, but it does not give any useful insight and it does not improve the fit. The same happens if we try, once the smoothing is the done, to consider the month as a nested effect within the beach, which we would have justified as being interchangeable measures once we isolate the seasonal component.

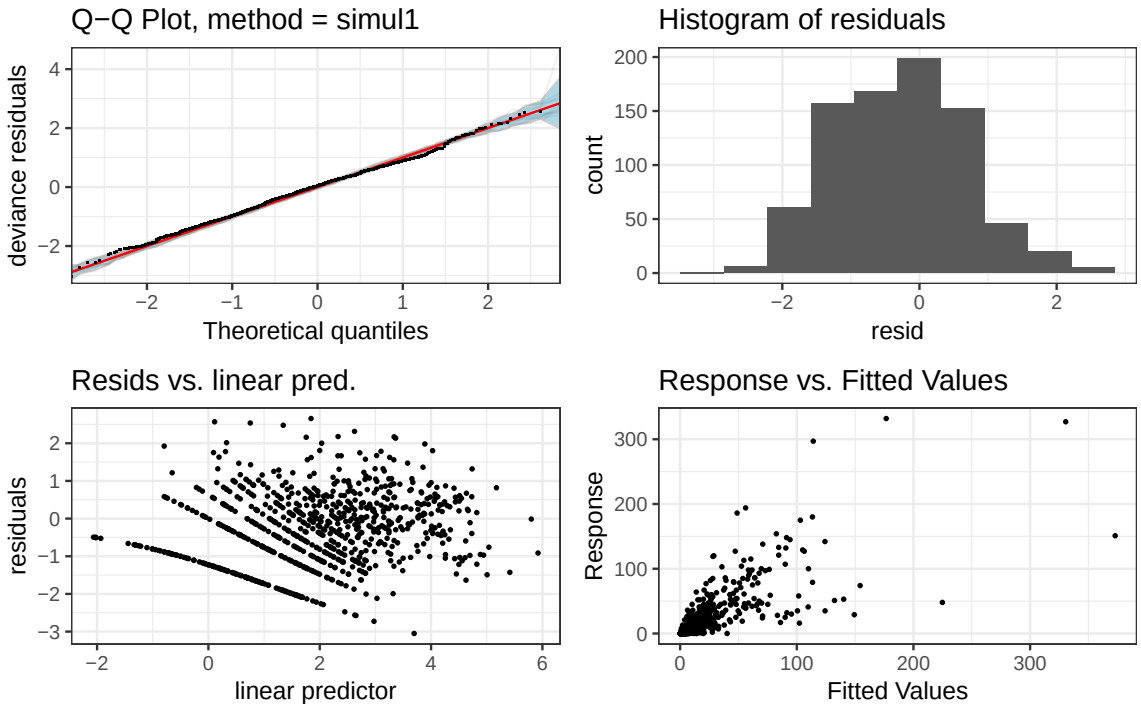


Figure 7: Diagnostics for our final model. A Negative Binomial GAM with beaches as random effects, cyclical smoothing on the month, smoothing on the inverse of the length of the beach and the geographical Region as fixed effect.

Then, we check whether our choice of using the length of the beach as an offset is appropriate, or if it would be better to include it in the model by smoothing on it. Thus, we test two additional models, the first smoothing on the length l and the second on the inverse of the length $1/l$. The AIC supports the latter one (5139.9) over the former and the model with l as offset (AIC $= 5140.1, 5142.5$ respectively).

We now test for the importance of the parameter 'type', which was always dropped by stepwise model selection in the GLMs. The deviance of the model including this covariate is 809.16 and it increases to 809.39 when we drop it. Since the models are nested, we can use a Chi-Squared test on the difference of deviances, which yields a p-value of 0.52 ($\chi^2_2 = 0.108$). With this in mind, it is not significant to include the type of beach in the final model.

Finally, we test for the inclusion of the yearly random factor. Again, we use a Chi-Squared test on the difference of deviances, by estimating the degrees of freedom using the R routine `anova.gam`. This leads to $\chi^2_{1.71} = 0.455$, which has a p-value of 0.27 and suggest to keep the more parsimonious model, as the year is not significant.

Figure 7 shows some diagnostics from our final model. Top-left panel contains the Q-Q plot, where the deviance residuals are very close to the theoretical quantiles. The histogram of the residuals in the top-right and the plot of residuals against the linear predictor in the bottom-left show a better behaviour than the GLMs, while there still seems to be some unexplained dispersion. In fact 18% of the residuals lie outside the $[-2, 2]$ interval. This is confirmed by the bottom-right panel, where we see that the higher the fitted value and the response are, the more likely it is to get a wrong fit. Nonetheless, this could just be a consequence of the distribution assumptions, as extreme values are more variables than small ones.

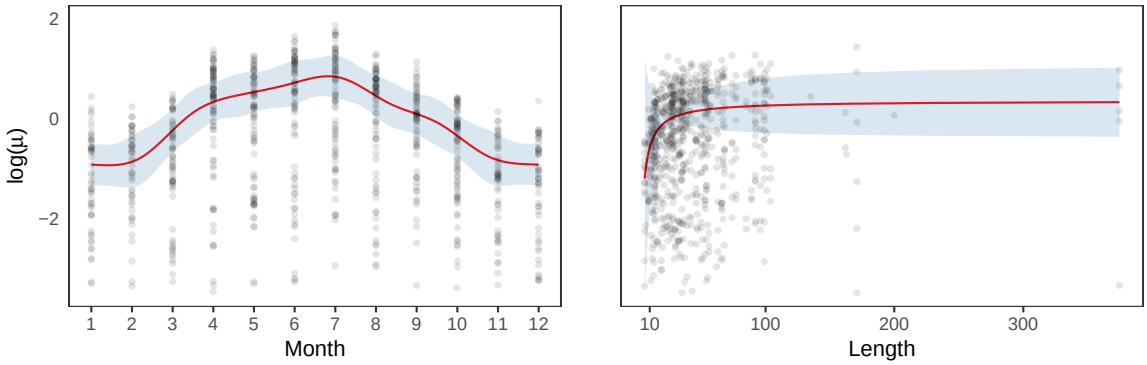


Figure 8: Effect on the log-mean of the smoothing functions on the month (left) and on the inverse of the length (right), as estimated by the Negative Binomial GAM.
The solid line shows the mean of the smoother and the transparent-blue surface show the 95% pointwise confidence interval, and the grey dots on the background are the GAM residuals. Both pictures only show the y-axes region sensible to the confidence intervals of the smoothers, but some residuals fall out of the range.

Overall, this final model seems to give an appropriate estimation of the actual distribution.

4 Discussion

In this section we delve deeper inside the model we got at the end of Section 3. In particular, we suppose that the responses, i.e. the number of cigarette butts, are distributed in the following way

$$\begin{aligned} y_i|\epsilon_i \sim \text{Poiss}(\mu_i\epsilon_i),\ \epsilon_i \sim \Gamma(\nu), \\ \log(\mu_i) = \sigma_m(m_i) + \sigma_l(1/l_i) + X\beta_{\text{Region}} + Zb_{\text{Beach}}, \end{aligned} \tag{4}$$

where ν is the parameter shape of the Gamma distribution; $\sigma_m(.)$ and $\sigma_l(.)$ are the smoothing functions for the month m and the inverse of the beach length l, respectively; $X\beta_{\text{Region}}$ is the region fixed effect and Zb_{Beach} is the beach random effect. We do not include any intercept, as we want to model the mean of each Region on its own.

We note that the ν parameter, giving the scale of the overdispersion factor ϵ in (4), has an estimate of 1.359.

	edf	Ref.df	Chi.sq	p-value
σ_l	1.00	1.00	5.77	0.02
σ_m	5.98	8.00	682.87	0.00
Beach (r.e.)	86.61	99.00	953.49	0.00

Table 3: Approximate significance of smooth terms. Note that $\sigma_m(.)$ and $\sigma_l(.)$ are the smoothing functions for the month and the last line gives the estimate for the random effect of the beach.

Table 3 gives us the approximate significance of the smoothing terms and, in particular, it points out that they are all significant. Figure 8 shows the smoothing functions on the month σ_m and on the inverse of the length σ_l, in the left and right panel respectively, as estimated by the Negative Binomial GAM. We see that seasonality influence the log-mean of the Poisson by giving a peak in summer, namely in July, and a minimum in winter. This is explained by the affluence of people to the beaches, which is strictly linked to the weather, which in Switzerland is mild in spring and

summer and cold in autumn and winter. More precisely, the more likely the beach is visited, the more likely we will find litter on it.

The smoothing effect of the length, computed through its inverse, points out that the smaller a beach is, the fewer the expected number of the cigarette butts is; nevertheless, the opposite does not hold. In fact, the right panel of Figure 8 shows that, after a certain length, the log-mean stabilises. The same happens with the estimate of the 95% CI. The plot shows that it is very tight for average-length beaches, while it widens for shorter and longer ones, but do not increases linearly with the length. This behaviour is a consequence of the choice of the inverse as representative of the length in the model, which increases the gap between small beaches, while reducing that between large ones. Nonetheless, this partially contradicts the hypothesis that pushed us to propose the length as an offset, namely that the dimension of the beach acts as a booster of the response. In fact, while it seems that small beaches do behave as supposed, the same does not hold for greater ones, where the influence of length seems to fade.

Region	Estimate	Std. Error	z value	$\Pr(>\|z\|)$
Graubunden	−0.20	0.74	−0.26	0.79
Eastern Switzerland	2.15	0.30	7.27	0.00
Basel	3.19	0.40	7.91	0.00
Bern	1.11	0.28	3.91	0.00
Zurich	2.25	0.19	12.16	0.00
Romandy	1.93	0.56	3.44	0.00
Central Switzerland	1.21	0.36	3.36	0.00

Table 4: Coefficients for the region fixed effect, as estimated in our final GAM. In last column all the p-values smaller than 0.001 have been rounded to zero.

Table 4 shows the estimates for the coefficient of the Region fixed effect. Basel is the region with the highest expected number of cigarette butts, with a coefficient of 3.19, followed by Zurich and Eastern Switzerland, with 2.25 and 2.15 respectively. Basel and Zurich are highly urbanized regions, which can explain such a high mean. In fact, 52 amongst the 58 beaches from Basel are within a city, the others being in the country, and Zurich counts 259 shores in the urban area (91 in city and 158 in the agglomeration) over a total of 358. The other regions differ only slightly, with the exception of Graubunden, which has the only non-significant estimate. This arises from the low variability within the region, as we can observe in Figure 3, which shows Graubunden on the far left.

Since we are satisfied with the fit and the explicability of our model, we move on to discuss its predictive abilities. In particular, we are interested in predicting the number of cigarette butts on the different beaches we studied for a certain day in August 2019. Since our model does not take into account the day, nor the year, the problem translates in predicting the number of cigarette butts for the month of August.

We will use two methods to derive confidence intervals for these predictions. The first one is based on the unconditional analysis in slide 180 of [1] and will give pointwise confidence intervals. The second one is done by simulation, as described in slide 184 of [1] and in the blog from [3] and will give simultaneous confidence intervals. The two methods take into account the uncertainty from fitting smoothing functions (σ_m, σ_l) and usual variability in the generalized mixed linear models.

Both methods will estimate a confidence interval for $\log(\mu)$, and then transform the confidence interval to the response level, using the inverse of the link function.

Figure 9 shows the results for the analytical method, which are very similar to those obtained by simulation. Table 5 shows the predictions for the Romandy region, around Lake Leman.

Beach ID	lower	upper
arve_carouge_battistellak	5.50	21.94
laclman_gland_kubela	1.46	12.20
Laclman_Gland_LecoanetS	4.49	22.6
laclman_vidy_santie	30.73	88.06
Ognonnaz-1	0.91	20.90

Table 5: Prediction for the number of cigarette butts to be found in August 2019 on beaches around the lake Leman (Romandy) with 95% confidence intervals.

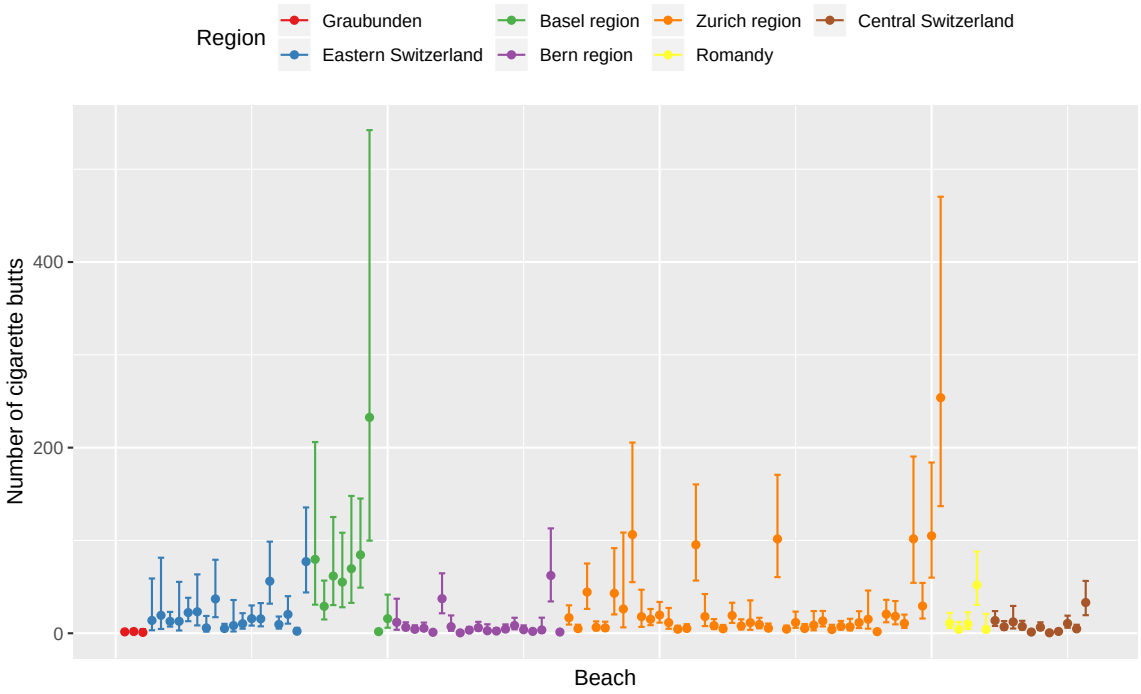


Figure 9: Predicted number of cigarette butts per Beach and 95% confidence intervals based on unconditional analysis of the residuals.

5 Conclusions

We fitted a Generalized Additive Model with Negative Binomial distribution to cigarette butts count on different shores in Switzerland. We regress their expected mean by smoothing on both the months and the inverse of the length of the beach, accounting for a fixed effect given by the Region and a random effect from the beach. We discover that seasonality highly affects the litter count, with a peak in summer and a minimum in winter. Furthermore, some regions have significantly higher means, often corresponding to more urbanized areas. The overall performance of the model is good, although it still leaves some variability unexplained.

Further improvements could consist in a more in detail study of the left out covariates such as the number of people visiting the beach, the percentage of the shore washed by the waves, the area of the beach.. Each one should be examined to determine if some data are missing at random or if they present a pattern. This would be needed in order to not lose too many observations and worsen

the fit of the models. As a way to deal with overdispersion, we mostly looked at quasi-likelihood method and parametric modeling, but did not have time to consider the Conway-Maxwell-Poisson distribution. One could also look at possible transformations of the data in order to stabilize the variance, at the risk of loosing the interpretability of the model.

References

[1] Anthony Davison. Modern regression methods, lecture notes, 2019.

[2] Hammerdirt. Beach litter, 2019. URL `https://mwshovel.pythonanywhere.com/dirt/beach_litter.html`.

[3] Gavin Simpson. Simultaneous intervals for smooths revisited, Dec 2016. URL `https://www.fromthebottomoftheheap.net/2016/12/15/simultaneous-interval-revisited/`.