REPORT DOCUMENTATION PAGE

Form Approved OMB NO. 0704-0188

1. REPORT DATE (DD-MM-YYYY)	2. REPORT TYPE	3. DATES COVERED (From - To)
05-01-2016	Final Report	23-Aug-2014 - 22-May-2015

4. TITLE AND SUBTITLE
Final Report: Sparse Event Modeling with Hierarchical Bayesian Kernel Methods

5a. CONTRACT NUMBER
W911NF-14-1-0488

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER
611102

6. AUTHORS
Kash Barker

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAMES AND ADDRESSES
University of Oklahoma
201 Stephenson Parkway
Five Partners Place, Suite 3100
Norman, OK 73019 -9705

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS (ES)
U.S. Army Research Office
P.O. Box 12211
Research Triangle Park, NC 27709-2211

10. SPONSOR/MONITOR'S ACRONYM(S)
ARO

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
65848-MA-II.2

12. DISTRIBUTION AVAILIBILITY STATEMENT
Approved for Public Release; Distribution Unlimited

13. SUPPLEMENTARY NOTES
The views, opinions and/or findings contained in this report are those of the author(s) and should not contrued as an official Department of the Army position, policy or decision, unless so designated by other documentation.

14. ABSTRACT
The research objective of this proposal was to develop a predictive Bayesian kernel approach to model count data based on several predictive variables. Such an approach, which we refer to as the Poisson Bayesian kernel model, is able to model the rate of occurrence of events (and subsequently, their likelihood of occurrence) based on historical evidence of the counts of previous event occurrences. The novel Bayesian kernel methods made use of: (i) the Bayesian property of improving predictive accuracy as data are dynamically obtained, and (ii) the kernel function which adds specificity to the model and can make nonlinear data more manageable. Early results show that the

15. SUBJECT TERMS
Bayesian statistics, kernel methods,

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	15. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Kash Barker
UU	UU	UU	UU		19b. TELEPHONE NUMBER 405-325-3721

Report Title

Final Report: Sparse Event Modeling with Hierarchical Bayesian Kernel Methods

ABSTRACT

The research objective of this proposal was to develop a predictive Bayesian kernel approach to model count data based on several predictive variables. Such an approach, which we refer to as the Poisson Bayesian kernel model, is able to model the rate of occurrence of events (and subsequently, their likelihood of occurrence) based on historical evidence of the counts of previous event occurrences. The novel Bayesian kernel methods made use of: (i) the Bayesian property of improving predictive accuracy as data are dynamically obtained, and (ii) the kernel function which adds specificity to the model and can make nonlinear data more manageable. Early results show that the Poisson Bayesian kernel model is more effective than the Poisson generalized linear model at modeling rates of occurrence especially for small data sets where regression-based methods often fail. The ability to model sparse data sets represents a positive step in modeling low-likelihood events often encountered in risk analysis.

Enter List of papers submitted or published that acknowledge ARO support from the start of the project to the date of this printing. List the papers, including journal references, in the following categories:

(a) Papers published in peer-reviewed journals (N/A for none)

Received Paper

TOTAL:

Number of Papers published in peer-reviewed journals:

(b) Papers published in non-peer-reviewed journals (N/A for none)

Received Paper

TOTAL:

Number of Papers published in non peer-reviewed journals:

(c) Presentations

Baroud, H. and K. Barker. 2014. Hierarchical Bayesian Kernel Models Applied to Event Data. INFORMS Annual Meeting, San Francisco, CA.

Number of Presentations: 1.00

Non Peer-Reviewed Conference Proceeding publications (other than abstracts):

Received	Paper

TOTAL:

Number of Non Peer-Reviewed Conference Proceeding publications (other than abstracts):

Peer-Reviewed Conference Proceeding publications (other than abstracts):

Received	Paper

TOTAL:

Number of Peer-Reviewed Conference Proceeding publications (other than abstracts):

(d) Manuscripts

Received	Paper	
01/05/2016	1.00	Hiba Baroud, Kash Barker. Poisson Bayesian Kernel Methods for Modeling Count Data, Computational Statistics and Data Analysis (04 2016)

TOTAL: **1**

Number of Manuscripts:

Books

Received Book

TOTAL:

Received Book Chapter

TOTAL:

Patents Submitted

Patents Awarded

Awards

Graduate Students

NAME	PERCENT_SUPPORTED	Discipline
Hiba Baroud	1.00	
FTE Equivalent:	**1.00**	
Total Number:	**1**	

Names of Post Doctorates

NAME	PERCENT_SUPPORTED
FTE Equivalent:	
Total Number:	

Names of Faculty Supported

NAME	PERCENT_SUPPORTED	National Academy Member
Kash Barker	0.11	
FTE Equivalent:	**0.11**	
Total Number:	**1**	

Names of Under Graduate students supported

NAME	PERCENT_SUPPORTED
FTE Equivalent:	
Total Number:	

Student Metrics

This section only applies to graduating undergraduates supported by this agreement in this reporting period

The number of undergraduates funded by this agreement who graduated during this period: 0.00

The number of undergraduates funded by this agreement who graduated during this period with a degree in science, mathematics, engineering, or technology fields:..... 0.00

The number of undergraduates funded by your agreement who graduated during this period and will continue to pursue a graduate or Ph.D. degree in science, mathematics, engineering, or technology fields:..... 0.00

Number of graduating undergraduates who achieved a 3.5 GPA to 4.0 (4.0 max scale):..... 0.00

Number of graduating undergraduates funded by a DoD funded Center of Excellence grant for Education, Research and Engineering:..... 0.00

The number of undergraduates funded by your agreement who graduated during this period and intend to work for the Department of Defense 0.00

The number of undergraduates funded by your agreement who graduated during this period and will receive scholarships or fellowships for further studies in science, mathematics, engineering or technology fields: 0.00

Names of Personnel receiving masters degrees

NAME
Total Number:

Names of personnel receiving PHDs

NAME	
Hiba Baroud	
Total Number:	**1**

Names of other research staff

NAME	PERCENT_SUPPORTED
FTE Equivalent:	
Total Number:	

Sub Contractors (DD882)

Inventions (DD882)

Scientific Progress

Technology Transfer

Sparse Event Modeling with Hierarchical Bayesian Kernel Methods

Proposal No. 65848MAII
Kash Barker (PI), University of Oklahoma

Background

One of the classical approaches used to analyze count data are Generalized Linear Models (GLM) [Agresti 2002, Cameron and Trivedi 2013]. The Poisson GLM is most commonly used to model count data. The method assumes that the rate to be estimated has an exponential relationship with a set of covariates representing coefficients for the different attributes, shown in Eq. (1). Under the Poisson GLM, the response follows a Poisson distribution, Eq. (2), and the log function is the link function that relates the set of covariates and coefficients to the response variable.

$$\hat{\lambda} = e^{\beta_i X} \tag{1}$$

$$P(y) = \frac{\lambda^y e^{-\lambda}}{y!} \tag{2}$$

Another type of GLM for modeling count data is the Negative Binomial GLM which relaxes the constraints of homoscedasticity imposed by the Poisson GLM [Cox 1983]. The Negative Binomial GLM assumes that the marginal distribution of the response follows a Negative Binomial distribution, Eq. (3), where k is the overdispersion parameter and λ is assumed to follow a Gamma distribution. The Negative Binomial GLM also assumes a log function for the link function and as a result, the response variable has an exponential relationship with the covariates.

$$P(y) = \frac{\Gamma\left(y + \frac{1}{k}\right)}{\Gamma(y+1)\Gamma\left(\frac{1}{k}\right)} \left(\frac{k\lambda}{1+k\lambda}\right)^y \left(\frac{1}{1+k\lambda}\right)^{\frac{1}{k}} \tag{3}$$

Approach

The Poisson Bayesian kernel model developed in this research is simple enough to avoid expensive computations but detailed enough to overcome issues in basic Bayesian modeling approaches, such as the Gamma conjugate prior, and in count data regression models, such as the GLM.

Poisson Bayesian kernel methods estimate the rate of occurrence of the event rather than estimating a deterministic value for the number of times the event is estimated to occur. A common distribution to model count data within a Bayesian framework is the Gamma-Poisson conjugate prior. The development

of the Poisson Bayesian kernel method discussed can be found in Floyd et al. [2014] and Baroud and Barker [2016]. The approach uses the Gamma conjugate prior as the basis of the model.

It is assumed that the parameter to be estimated is the rate of occurrence, $\lambda > 0$, which follows a Gamma prior distribution with parameters $\alpha > 0$ and $\beta > 0$, as shown in Eq. (4). For the likelihood function, the product of the Poisson density function, shown in Eq. (5), is used, since this is a Gamma-Poisson conjugate prior approach.

$$P(\lambda) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} \lambda^{\alpha-1} \mathrm{e}^{(-\beta\lambda)} \tag{4}$$

$$L = \prod_{i=1}^{m} P(y_i) = \prod_{i=1}^{m} \frac{(\lambda_i^{y_i} e^{-\lambda_i})}{y_i!} = \frac{\lambda_i^{\sum_{i=1}^{m} y_i} e^{-m\lambda_i}}{\prod_{i=1}^{m} y_i!} \tag{5}$$

Rearranging the product of the likelihood function and the prior distribution function results in a Gamma posterior distribution where $\alpha^* = \sum_{i=1}^{m} x_i + \alpha$ and $\beta^* = m + \beta$.

$$\begin{aligned} P(\lambda|x) &= \left(\frac{\beta^{\alpha}}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda}\right)\left(\lambda^{\sum_{i=1}^{m} y_i} e^{-m\lambda}\right) \\ &= \frac{\lambda^{(\sum_{i=1}^{m} y_i + \alpha - 1)}\, e^{-\lambda(m+\beta)} (n+\beta)^{\sum_{i=1}^{m} y_i + \alpha}}{\Gamma(\sum_{i=1}^{m} y_i + \alpha)} \\ &= \text{Gamma}\left(\alpha^*, \beta^*\right) \end{aligned} \tag{6}$$

This result is the basic Gamma conjugate prior approach used in Bayesian analysis. This approach assumes the notion of exchangeability meaning that for different sets of training and testing data, the resulting posterior parameter will be similar since they are a function of the prior parameter, the size of the dataset, and the summation of all the data points. The characteristics of each outcome are not taken into consideration in this case, but rather the overall property of the dataset [Mackenzie et al. 2014].

The Poisson Bayesian kernel approach extends the notion of the conjugate prior such that the posterior parameters computation not only depends on the prior parameters and the historical data but also on the attributes through the kernel matrix. The parameters for the Bayesian kernel model for counts are expressed in Eqs. (7) and (8). **K** is the $m \times m$ kernel matrix, **Y** is an $m \times 1$ vector containing the output data associated with the m observations of **X**, and **V** is an $m \times 1$ vector containing ones. Each entry in the kernel matrix represents the similarity measure between the attributes of the testing set and the training set. As such, the new data point is compared with the training set and according to the similarities of the attributes, new values for

the parameter of the posterior distribution are computed. Note that in this case, the training and testing sets are assumed to have the same size, m. However, when the model is deployed, the sets can be of different sizes, and in some cases, the testing set could include only one data point such as in a leave-one-out analysis that will be illustrated in the case study.

$$\alpha^* = \mathbf{KY} + \alpha \tag{7}$$

$$\beta^* = \mathbf{KV} + \beta \tag{8}$$

As with other statistical and mathematical models, there are a few assumptions underlying the deployment of such modeling approach. Even though the form of the prior distribution is known from the conjugate prior, the model user would still need to identify the values of the prior parameters. Oftentimes, the priors are either assumed to be known or are assigned such that the prior distribution is non informative. In other cases, these parameters are estimated using data and prior knowledge by matching the sample mean and variance to those of the prior distribution [MacKenzie et al 2014, Carlin and Louis 2008]. Further discussion on the choice and impact of prior parameters is provided in the case study of this chapter. Another assumption to consider is the choice of the kernel function which depends on the application and the model user. This research uses the most popular kernel function, the radial basis function (RBF) in Eq. (9), where $k(\mathbf{x}_i, \mathbf{x}_j)$ is one entry in the matrix $\mathbf{K}$ representing the kernel function between the attributes of the i^{th} and j^{th} data points.

$$k(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma^2}\right) \tag{9}$$

The rate for the new data point follows then a Gamma distribution with parameters α^* and β^*. As a point estimate for this parameter, the expected value of the posterior distribution is considered, shown in Eq. (10) as the ratio of the Gamma distribution parameters α^* and β^*. Note that a different point estimate for the rate can be used such as the median, the mode, or the variance, depending on the type of problem and the model users.

$$\hat{\lambda} = \frac{\alpha^*}{\beta^*} \tag{10}$$

Goodness of Fit Measures and Prediction Accuracy

To assess the performance of the model, goodness of fit measures are analyzed to identify the capability of the model to capture data patterns. The empirical analysis and the case study compare the Poisson Bayesian kernel (PBK) model

to other classical methods for modeling count data, the Poisson generalized linear model (GLM) and the Negative Binomial GLM. The Poisson and Negative Binomial GLM assume that the rate to be estimated has an exponential relationship with a set of covariates representing coefficients for the different attributes, $\hat{\lambda}_{PGLM} = e^{\beta_i X}$, while the predicted rate for the PBK is equal to the expected value of the posterior probability distribution, $\hat{\lambda}_{PBK} = \frac{\mathbf{KY}+\alpha}{\mathbf{KV}+\beta}$.

The functional values of two metrics are used to compare how well the models fit the data and are able to explain the variance. The first metric is the *deviance*, which computes the difference in the log-likelihood function between the fitted model and the saturated model, Eq. (11), where y_i is the true value of the data point and $\hat{\lambda}$ is the estimated rate for the particular data point.

$$D = 2 \times \left(l(\boldsymbol{y}|\boldsymbol{y}) - l\left(\hat{\boldsymbol{\lambda}}|\boldsymbol{y}\right)\right) \tag{11}$$

The second metric used is the functional value of the *log-likelihood*, shown in Eq. (12), which is to be maximized. The log-likelihood function represents the joint probability of the observed data as a function of the parameter of interest which is $\hat{\lambda}$ in this case. The larger the value of this function, the better the model is able to capture the data patterns using the estimated parameters.

$$l\left(\hat{\boldsymbol{\lambda}}|\boldsymbol{y}\right) = \sum_{i=1}^{m} [y_i \ln(\hat{\lambda}_i) - \hat{\lambda}_i - \ln(y_i!)] \tag{12}$$

The ultimate objective of building the Poisson Bayesian kernel model is to deploy it in risk analysis problems, such as predicting the frequency of disruptions in a particular network system. While the goodness of fit is important to assess whether the model is capturing the pattern and variability in the data, is it equally important to analyze the prediction power of a statistical model if it is going to be used for forecasting purposes. Prediction accuracy is assessed by the out-of-sample error, which accounts for the discrepancy between the estimated parameter and the actual observation of data points that were not in the set used to train the model.

Adding complexity to the model will decrease the training error but may cause overfitting at some point resulting in a poor prediction accuracy when the model is applied to an independent data set. To validate the prediction power of the models, several metrics are evaluated to assess the out-of-sample error, and they are summarized in Table 1.

The PBK is applied to several data sets, and its performance is compared to the Poisson and Negative Binomial GLM using the goodness of fit and prediction accuracy metrics discussed previously. Most of the data sets are

similar in terms of the number of predictors and the size of the data. One of the sets has a larger number of predictors for a small data set, and another is a large data set with a small number of predictors. Note that the number of predictors in all the models is held constant across the data sets to ensure consistency in the comparison. Also, the parameters of the prior distribution for all data sets are assumed to be $\alpha = \beta = 1$, also in order to maintain consistency in the evaluation of the performance of all models.

A holdout analysis is performed where each data set is randomly split into training and testing sets for 100 trials. Traditional holdout analyses would train the model on a portion of the data and deploy it on the testing set to make predictions and compare them to the actual observations. With the PBK, an intermediate step in the training process is added to tune the unknown parameter, σ, in the kernel function. This parameter is optimized based on the minimum mean square error. As a result, 30% of the data was used for testing the model, with 50% of the data used as a training set and 20% as a tuning set. The training and the tuning sets were then combined into one training set to perform the testing. For each of the three models, the estimated rate of occurrence is computed for the testing test and used to evaluate the deviance, the log-likelihood functional value, and the four out-of-sample error measurements given the observed values. This process is repeated 100 times where, at each iteration, random samples of training, tuning, and testing sets are chosen. Table 3 contains a summary of the analysis. The performance metrics values presented in the table below are the average values of the performance measures evaluated over 100 trials. PBK refers to the Poisson Bayesian kernel model and PGLM refers to the Poisson GLM, and NBGLM refers to the Negative Binomial GLM. Recall that the model with a smaller deviance and errors and a larger log-likelihood functional value is a better model.

Overall, there are five out of seven data sets for which the Poisson Bayesian kernel model outperforms the Poisson and Negative Binomial GLM in terms of the predictive accuracy. In particular, those five cases are all among the six small data sets. The RMSE, NRMSEM, NRMSED, and MAE all behave similarly for all the datasets and lead to the same conclusion of the model performance, except for a minor difference in the *Migration to Edinburgh* where the PBK performs similarly to the NBGLM and slightly worse than the Poisson GLM in terms of MAE values. With respect to goodness of fit measures, the GLMs perform better than the PBK. Overall, the Negative Binomial fits the best. PGLM and NBGLM perform similarly in the two data sets for which the GLM outperforms the PBK in the predictive accuracy, *Customer* and *Murder*. The Poisson Bayesian kernel model appears to be a good model for prediction

purposes when the data set is small with a small number of predictors, a situation known to cause issues with regression modeling.

Table 1: Prediction error measurement metrics.

Prediction accuracy metrics	Formula
Root Mean Square Error	$RMSE = \frac{1}{n}\sqrt{\sum_{i=1}^{n}\left(Y_i - \hat{\lambda}_i\right)^2}$
Normalized Root Mean Square Error	$NRMSEM = \frac{\frac{1}{n}\sqrt{\sum_{i=1}^{n}\left(Y_i - \hat{\lambda}_i\right)^2}}{Y_{maximum} - Y_{minimum}}$
Mean Absolute Error	$MAE = \frac{1}{n}\sum_{i=1}^{n}\left\|Y_i - \hat{\lambda}_i\right\|$

Table 2: Description of data sets in the Poisson Bayesian kernel model validation study.

Data set	Number of attributes	Data set size	Dependent variable	Predictors
Crime	4	50	Crime rate	Race, percentage of high school graduates, percentage below poverty level, percentage with a single parent
Murder	4	51	Murder rate	Race, percentage of high school graduates, percentage below poverty level, percentage with a single parent
Mussels	8	45	Number of species of mussels	Area, number of stepping stones (intermediate rivers) to 4 major species-source river systems, concentration of nitrate, solid residue, concentration of hydronium
Customer	5	110	Number of customers visiting a store from a particular region	Number of housing units in the region, average household income in the region, average housing unit age in the region, distance to the nearest competitor, distance to the store
West Nile virus in birds	4	46	Cases of virus in birds	Numbers of farms, area, population, human density
West Nile virus in equines	4	46	Cases of virus in equines	Numbers of farms, area, population, human density
Migration to Edinburgh	4	33	Number of apprentices migrating	Distance, population, degree of urbanization, direction from Edinburgh

Table 3: Performance metrics results for the empirical analysis.

Data	Metrics	PBK	PGLM	NBGLM
Crime	LL	-276.25	-256.56	**-155.61**
	DEV	352.87	**313.49**	343.14
	RMSE	**26.47**	33.15	37.69
	NRMSEM	**0.28**	0.35	0.39
	NRMSED	**0.89**	1.13	1.29
	MAE	**21.26**	21.97	23.19
Murder	LL	-120.96	**-77.79**	**-77.79**
	DEV	107.69	**21.38**	**21.38**
	RMSE	9.81	**3.85**	**3.86**
	NRMSEM	0.28	**0.17**	**0.17**
	NRMSED	0.98	**0.58**	**0.58**
	MAE	4.68	**2.59**	**2.59**
Mussels	LL	-97.33	**-78.91**	**-78.72**
	DEV	66.55	29.71	**26.43**
	RMSE	**5.60**	5.84	5.83
	NRMSEM	**0.27**	0.31	0.31
	NRMSED	**0.96**	1.08	1.07
	MAE	**4.00**	4.32	4.31
Customer	LL	-230.02	**-194.46**	**-194.46**
	DEV	149.15	**78.04**	**77.69**
	RMSE	5.13	**3.58**	**3.58**
	NRMSEM	0.18	**0.13**	**0.13**
	NRMSED	0.77	**0.55**	**0.55**
	MAE	3.78	**2.75**	**2.75**
West Nile virus in birds	LL	-135.35	-103.94	**-80.76**
	DEV	181.74	118.93	**36.14**
	RMSE	**7.78**	8.44	9.14
	NRMSEM	**28.85**	33.47	36.44
	NRMSED	**98.22**	113.65	123.84
	MAE	**4.91**	5.09	5.23
West Nile virus in equines	LL	-40.12	-40.47	**-39.42**
	DEV	43.42	44.12	**32.57**
	RMSE	**1.75**	2.08	2.05
	NRMSEM	**0.30**	0.41	0.41
	NRMSED	**0.95**	1.24	1.25
	MAE	**1.17**	1.33	1.29
Migration to Edinburgh	LL	-127.98	-106.03	**-64.46**
	DEV	442.9	146.62	**25.78**
	RMSE	**31.23**	32.29	32.45
	NRMSEM	**0.43**	0.53	0.51
	NRMSED	**1.32**	1.61	1.55
	MAE	15.98	**14.71**	15.86

Case Study

With over 200 lock chambers and more than $150 million worth of goods flowing yearly [US Army Corps of Engineers 2011], the inland waterway system plays an important role in the nation's economy. Unfortunately, the system's reliability is declining due to the aging components of the network [Grier 2009]. According to the American Society of Civil Engineers' most recent report card on America's infrastructure, inland waterways received a grade of "D^-" while dams received a grade of "D". Among the most common causes for the degrading status of inland waterways are aging components. On average, dams in the United States are 52 years old, and by the year 2020, 70% of the dams will be over 50 years old [ASCE report card 2013]. As a result, locks and dams are frequently closed for unscheduled or scheduled maintenance which causes delays in the flow of commodities and incurs large economic losses across the nation. In 2009, 90% of locks and dams in the US experienced service interruption resulting in an average of 52 delays a day.

The Poisson Bayesian kernel model is applied to analyze the frequency of lock closure due to disruptive events on the Mississippi River transportation network. The network has 29 locks acting as key connectors between different ports nationwide. The navigation system reflects 9,000 miles of navigable waterway with 70.5% of the U.S. inland waterway commodity flowing through the network.

The data, retrieved from the database collected by the U.S. Army Corps of Engineers [2011], contains detailed information on each lock's characteristics including the river mile, the total number of vessels passing by the lock, the total tonnage, and the frequency and average delay for the vessels and tows experiencing delay time due to the lock's closure. Data is available on the yearly frequency of closure for each lock which is considered in this case the outcome to be estimated. A sample of the data is represented in Table 4.

The goal of deploying the Bayesian kernel model is to obtain an accurate prediction of the frequency of disruptions to inform preparedness strategies and investment decision making. Using the Poisson Bayesian kernel model, decision makers are able to produce a probability distribution of the number of times a particular lock and dam will close each year. The distribution can be used to improve risk management along the inland waterways and make them a more reliable transportation system.

As a first step, the prediction accuracy of the PBK model is tested in comparison with the PGLM and the NBGLM. Similarly to the analysis done in the empirical study, a

holdout analysis is performed to assess the goodness of fit and prediction accuracy of PBK for the inland waterway, and the results are summarized in Table 5.

Table 4: Sample of the inland waterway disruption data.

	Y	X_1	X_2	X_3	X_4	…
Lock & Dam	Closure Frequency	River Mile	Vessels	Tonnage	Lockages	. . .
L&D 3	0	797	9,397	6,747	4,406	. . .
L&D 13	6	523	2,810	14,545	3,155	. . .
L&D 2	0	815	4,478	6,735	2,893	. . .
L&D 20	23	343	2,508	20,828	3,582	. . .
L&D 22	40	301	2,280	22,476	3,486	. . .
L&D 8	6	679	4,333	10,277	2,620	. . .
⋮	⋮	⋮	⋮	⋮	⋮	. . .

Table 5: Performance metrics results for the inland waterway data analysis.

	Metrics	**PBK**	**PGLM**	**NBGLM**
Full model	LL	-285.06	-148.06	**-75.09**
	DEV	486.03	211.65	**23.89**
	RMSE	**32.82**	63.94	131.03
	NRMSEM	**0.34**	0.66	1.24
	NRMSED	**0.94**	1.88	3.45
	MAE	**21.53**	33.23	57.15
Best model - PGLM	LL	-252.25	**-146.01**	
	DEV	420.53	**208.06**	
	RMSE	**32.60**	42.37	
	NRMSEM	**0.34**	0.46	
	NRMSED	**0.95**	1.34	
	MAE	**20.76**	25.08	
Best model - NBGLM	LL	-238.07		**-78.13**
	DEV	391.33		**24.15**
	RMSE	**28.46**		46.74
	NRMSEM	**0.30**		0.54
	NRMSED	**0.87**		1.56
	MAE	**18.00**		26.01

According to the values of the average out-of-sample error expressed in the four metrics, RMSE, NRMSEM, NRMSED, and MAE, PBK does a better job at making accurate predictions of the average frequency of lock and dam closures even though based on the values of the log-likelihood and deviance, GLM, more specifically the Negative Binomial GLM is better at fitting the data. One of the reasons a GLM might not be providing good prediction errors is overfitting. In order to check whether the results obtained, after fitting a full model that includes all covariates, are due to the GLM overfitting the data, the analysis is performed given the best version of the GLMs. The selection of covariates for each of the Poisson and Negative Binomial GLM is based on Akaike's Information Criterion (AIC) that penalizes additional parameters contributing to the model complexity. In terms of goodness of fit measures, both reduced GLMs did not express any change in the values of the log-likelihood and deviance from the full model. However, the prediction accuracy improved significantly for the reduced model with about 60% decrease in the values of RMSE, NRMSEM, NRMSED, and MAE for NBGLM and about 30% decrease for PGLM error measurement values. The covariates selected for the best models were also included in the PBK evaluation for consistency. The PBK still performed better than the best version of both GLMs and maintained better predictive error measures even though the reduced versions of GLMs significantly improved their prediction accuracy.

One of the advantages of using Bayesian methods in risk analysis is the flexibility of the approach in (i) establishing assumptions, and (ii) interpreting the results. Any prior belief about the risk measure to be estimated can be embedded in the prior distribution. Determining the prior parameters can be challenging and can result in significant implications on the posterior parameters' estimation. So far, the analysis considered the same prior distribution with prior parameters $\alpha = \beta = 1$ to insure consistency in the empirical study across the different data sets and models. This section examines the implications of changing the priors on the posterior parameters.

In risk analysis problems, experts in the field can help in assessing any prior knowledge about the parameter to be estimated. Ideally, risk managers are interviewed, and using probability elicitation techniques, a prior probability distribution is defined. Three levels of knowledge are considered in this case that influence the estimation of the priors. For each case scenario, the posterior frequency of disruptions is computed and compared to results from fitting a PGLM and a NBGLM. The distribution of the RMSE across the three models under each case scenario is used to assess the impact of the priors.

The first approach assumes the experts have a perfect knowledge about the frequency of disruptions and the prior parameters are estimated from the data using the method of moments, Eq. (13) , where $\bar{Y}$ and s^2 are respectively the mean and variance of the historical data.

$$\alpha = \frac{\bar{Y}^2}{s^2}$$
$$\beta = \frac{\bar{Y}}{s^2} \tag{13}$$

The plot in Figure 1 shows that the distribution of RMSE values is skewed towards the smaller values (around 25), while the PGLM and NBGLM distributions of RMSE values are spread across larger values with thicker tails. The dashed lines correspond to the mean RMSE showing that PBK performs the best in terms of prediction accuracy.

The second case scenario assumes that the risk managers have some prior knowledge but it is not perfect like in the first approach. The bias introduced by the risk managers is modeled with a random noise and the distribution of RMSE for the three models is depicted in Figure 2. The expected value of RMSE and the overall distribution of the values are both quite similar to the case where the knowledge is assumed to be perfect. Note that the added noise in this case is not very significant. If the risk manager expressed a stronger bias, more noise would be added which could impact the prior parameters estimation and ultimately the posterior distribution and predicted values of the frequency of disruption.

For the third approach, it is assumed that the risk managers have no prior knowledge, or there is no access to a reliable source of information to estimate informed prior parameters. Therefore, the priors are arbitrarily determined and the distribution of the RMSE values is plotted in Figure 3. The smaller values of RMSE still have the highest frequency in the distribution of RMSE for PBK; however this peak is now centered around values equal to 50 as opposed to 25 in the first two approaches. The overall distribution shifted to the right, towards larger values of the RMSE, and the distribution is overlapping with PGLM and NBGLM RMSE distributions. In addition, the expected value of RMSE for PBK increased and is approaching the RMSE expected value of the PGLM, although it is still significantly smaller.

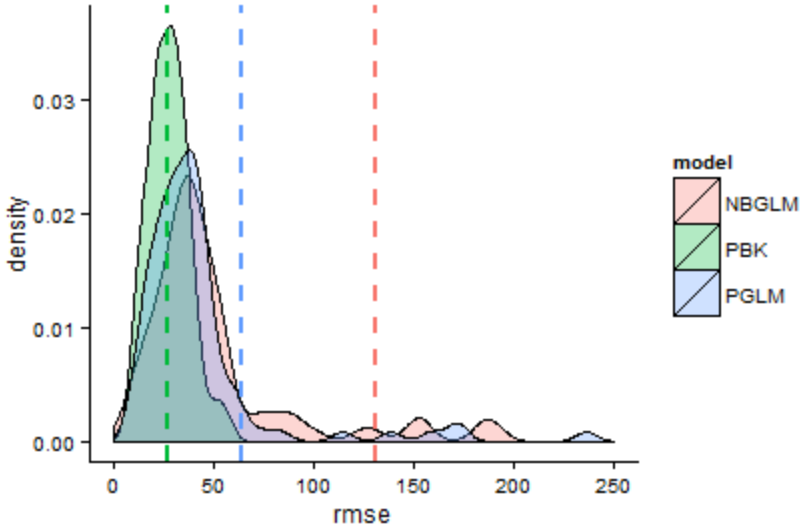

Figure 1: RMSE distribution with perfect prior knowledge.

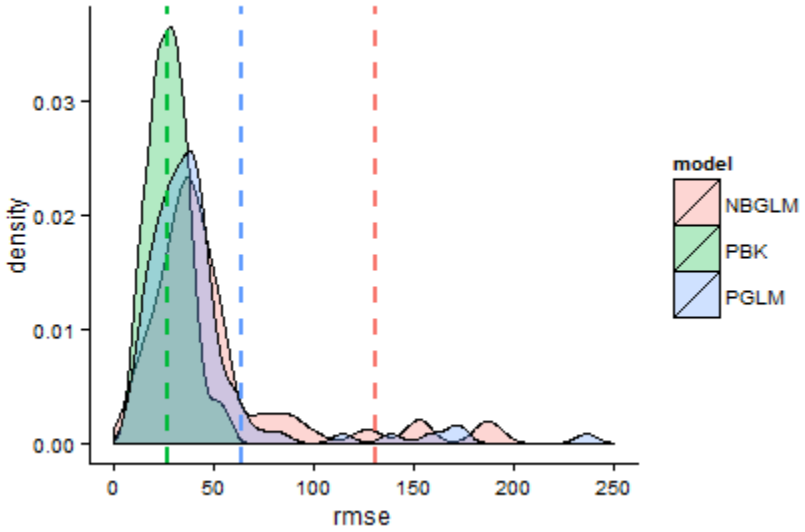

Figure 2: RMSE distribution with imperfect prior knowledge.

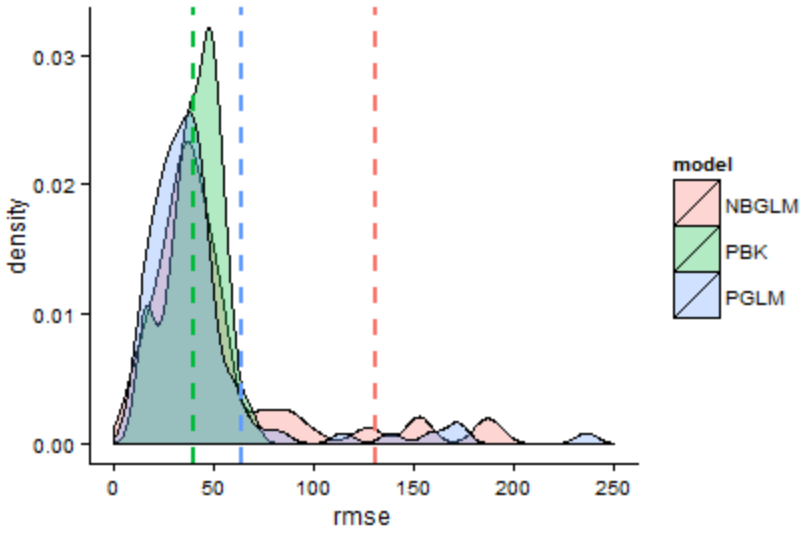

Figure 3: RMSE distribution with no prior knowledge.

The selection of prior parameters has an implication on the form of the posterior distribution. A poorly formulated prior distribution can impact the performance of the PBK in predicting the frequency of lock and dam closures. On the other hand, a perfect prior distribution relying solely on the historical data is unrealistic. Therefore, the model user must carefully formulate the prior distribution and prior parameters to ensure accurate posterior inferences.

The model developed in this chapter can impact the decision making process for the protection and rehabilitation of the U.S. inland waterways. As mentioned earlier, this critical infrastructure system is suffering from aging components resulting in frequent disruptions of the flow of commodities across the nation. The Department of Homeland Security announced a set of grant programs to protect and rehabilitate critical infrastructure systems. These grants are normally assigned based on the priority of the rehabilitation project due to the limited availability of resources. Using the hold-one-out analysis approach, the PBK is used to produce a rank of the locks and dams of the inland waterway network. Such a ranking of an infrastructure system's components is one way to implement data-driven risk analysis into real world decision making. Table 6 contains the top five locks & dams with the highest predicted frequency of closures per year.

Table 6: Locks/dams with highest frequency of closures

Ranking	Lock/Dam ID
1	L&D 27
2	Mel Price L&D

3	L&D 19
4	L&D 21
5	L&D 22

Frequent disruptions might be one indication of a component's reliability and urgent need for rehabilitation. As such, the ranking produced in the table above can either be used to allocate grants accordingly or it can be integrated into a multiobjective decision tool that incorporates other factors into the assessment of a rehabilitation project.

Conclusions

Bayesian kernel methods are powerful tools in forecasting data. These models make use of the Bayesian property by relying on historical data and experts' knowledge, but they also add more specificity to the model by using the kernel function. Gaussian Bayesian kernel models became very popular recently and were extended and applied to a number of classification problems. An important extension to those models is the non-Gaussian model which gives more flexibility in applying this methodology to all types of data set, however, there has been no Bayesian kernel model in the literature that addresses count data.

This chapter introduced count data modeling to the class of Bayesian kernel methods. Using the notion of the conjugate prior, the rate of occurrence is assumed to follow a Gamma prior and posterior distribution using the Poisson likelihood function. The parameters of the posterior distribution are constructed using results from the classical Bayesian Gamma conjugate prior and the exchangeability argument.

The Poisson Bayesian kernel model presented in this chapter is empirically tested and compared with the classical Poisson and Negative Binomial GLM. The three models were used to fit several datasets having similar characteristics in terms of the size of the data and the number of predictors. The evaluation of the performance of each model is based on the values of metrics corresponding to the goodness of fit and prediction accuracy. Based on the results obtained, the Poisson Bayesian kernel model outperforms the Poisson and Negative Binomial GLM in the majority of the sets for most of the performance metrics representing the out-of-sample error. Also, the Poisson Bayesian kernel model is potentially a better model for small-sized data sets having few predictors. Such a result can be very useful in risk analysis applications to estimate the rate of occurrence of a certain disruption in transportation systems or power grids. In such cases, data can be limited due to the lack of occurrence of the event and the possible factors that might cause a disruption. The need for a more accurate estimation of the rate of disruption can help save lives and lead to more efficient preparedness and recovery investment and allocation.

The Poisson Bayesian kernel model is illustrated using waterway transportation network data of the frequency of lock closure along the Mississippi river, and compared to the classical Poisson and Negative Binomial GLM for the six metrics used in the empirical study. While GLMs exhibit a better fit of the data, the Bayesian kernel model produces a smaller out-of-sample error suggesting a better prediction power. Accurate predictions of the frequency of disruptions are used to rank the locks and dams and allocate rehabilitation resources accordingly. Realistically, the rank would be one of many criteria used in the decision making process. This chapter addresses the prediction of risk of infrastructure disruptions, the second step would be to understand and quantify the interdependent economic impacts of a disruptive event and how they influence preparedness decision making accordingly.

References

Agresti, A. (2002). *Categorical Data Analysis*, 2nd edition. Hoboken, NJ: Wiley-Interscience.

American Society of Civil Engineers. (2013b). Report Card for America's Infrastructure 2013.

Baroud, H. & Barker, K. (2016). Poisson Bayesian Kernel Methods for Modeling Count Data. To be submitted to *Computational Statistics and Data Analysis*.

Cameron, A. C., & Trivedi, P. K. (2013). *Regression Analysis of Count Data* (Vol. 53). Cambridge university press.

Carlin, B. P., & Louis, T. A. (2008). *Bayesian Methods for Data Analysis*. CRC Press.

Cox, D. R. (1983). Some remarks on overdispersion. *Biometrika*, *70*(1), 269-274.

Floyd, M. S., Baroud, H., & Barker, K. (2014). Empirical analysis of Bayesian kernel methods for modeling count data. In *Systems and Information Engineering Design Symposium (SIEDS), 2014* (pp. 328-333). IEEE.

Grier, D. V. (2009). The declining reliability of the US inland waterway system. *Institute for Water Resources*, Army Corps of Engineers. Alexandria, VA.

MacKenzie, C. A., Trafalis, T. B., & Barker, K. (2014b). A Bayesian Beta kernel model for binary classification and online learning problems. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, *7*(6), 434-449.

U.S. Army Corps of Engineers. (2011). Interactive access to website. http://www.ndc.iwr.usace.army.mil//lpms/lpms.htm.