Paper 8441-2016

exSPLINE That: Explaining Geographic Variation in Insurance Pricing

Carol Frigo and Kelsey Osterloo, State Farm Insurance

ABSTRACT

Generalized linear models (GLMs) are commonly used to model rating factors in insurance pricing. The integration of territory rating and geospatial variables poses a unique challenge to the traditional GLM approach. Generalized additive models (GAMs) offer a flexible alternative based on GLM principles, with a relaxation of the linearity assumption. We will explore two approaches for incorporating geospatial data in a pricing model using a GAM based framework. The ability to incorporate new geospatial data and improve traditional approaches to territory ratemaking results in further market segmentation and a better match of price to risk.

Our discussion highlights the use of the high-performance GAMPL procedure, which is new in SAS/STAT® 14.1 software. With PROC GAMPL, we can incorporate the geographic effects of geospatial variables on target loss outcomes. We will illustrate two approaches. In our first approach, we begin by modeling the predictors as regressors in a GLM, and subsequently model the residuals as part of a GAM based on location coordinates. In our second approach, we model all inputs as covariates within a single GAM. Our discussion will compare the two approaches and demonstrate visualization of model outputs.

INTRODUCTION

The practice of rating policies based on their geographic territory is a part of most insurance rating plans. Traditionally, territories have been built based on government boundaries such as county, city and zip code. Since these territories are large geographic areas, much of the high level variation can be explained, but there is still more information that can be extracted to account for the deeper granular differences between policyholders. The current approach looks at territories as spatial grid boundaries; that is, as square grid cells based on latitude and longitude ranges. To model the observed losses of these grid cells, we currently credibility weight each cell with its neighboring grid cells. Our proposed methods explore generalized additive models (GAMs) to capture the multi-directional variation and incorporate new location variables to better address the individuality of each policy.

GENERALIZED ADDITIVE MODELS

Generalized linear models (GLMs) relax the assumptions of general linear models by allowing for link functions that transform the mean of the response variable, and distributions other than normal. GAMs[1] go one step further by relaxing the linearity assumption to allow smoothing functions. In a GAM, the response can depend on both parametric and nonparametric predictors. The parametric predictors can be linear effects involving continuous and classification variables, as in GLMs. The nonparametric predictors can be smooth functions of one or more continuous variables. GAMs can be expressed as follows:

$$g(u) = f_0 + f_1(x_1) + f_2(x_2, x_3) + \cdots + f_p(x_p)$$

where u = E(Y), g(.) is a link function applied to the mean, and f_p (.) characterizes the linear effect or the dependency of the smoothing function in the neighborhood of x_p. The smoothing functions can take a variety of forms, such as local average, running average, running line, kernel-based, and spline.

The increased flexibility of the GAM allows for better fits of the contoured data that are not easily parameterized in a linear or higher degree polynomial model. The spline terms can be univariate or multivariate, where the multivariate splines include the interaction of the variables, thus capturing more of the joint variation in the data.

GAM MODELING IN SAS – THE GAMPL PROCEDURE

GAMPL, which stands for Generalized Additive Model by Penalized Likelihood estimation, is a new high-performance procedure in the SAS/STAT® 14.1 software[2]. This is not simply the high-performance version of PROC GAM – the two procedures have different underlying regression equations. PROC GAM uses smoothing splines and is considered better for smaller amounts of data with known degrees of freedom. The GAMPL procedure uses regression splines and is better for searching for the optimal degrees of freedom and automatic model building. In the procedure's MODEL statement, each regressor must be enclosed within SPLINE() or PARAM(). SPLINE( ) is used to specify spline functions of continuous regressors, and PARAM() is used to specify linear effects of continuous or classification regressors. To begin with a simple example, we can model latitude and longitude against pure premium (dollars of loss per policy, aggregated to the grid cell level). Latitude and longitude values are rounded to the hundredths place, effectively creating .01°x.01° (or 1km by 1km) square grid cells. Our dataset contains over 600,000 observations, each being a unique latitude/longitude combination, covering the entire state. The modeling code for this example is:

```
PROC GAMPL data=StateX_Crime;
        MODEL PurePrem = spline(Latitude Longitude / maxdf=200 maxknots=20000);
        OUTPUT out=StateX_output    pred= predicted;
RUN;
```

Here, we have a bivariate spline constructed by latitude and longitude, creating a smoothed rating surface topology over the geographic space. The procedure's default maximum degrees of freedom for a bivariate spline term is 20 (the default for univariate splines is 10), and the default maximum number of knots is 2,000. Due to the complexity of the data, we need more than the default degrees of freedom and knots to appropriately capture the variation, and thus have expanded the search area, requesting the optimal degrees of freedom selected to be between one and two hundred and the optimal number of knots to be at most 20,000.

The degrees of freedom (DFs) are a measure of nonlinearity, and can be non-integer due to the effects of the spline. DFs are reported both for the model as a whole, and individually for each spline term. Smaller model DFs imply a simpler, easier to interpret model. Spline DFs close to one indicate that smoothing is not needed – the variable has a linear relationship with the target and should instead be included as a parameter. The formula found in the Details section[3] of the documentation for the GAMPL procedure provides an intuitive way to parameterize the complexity of the spline model. The formula shows the natural extension of the definition of degrees of freedom for linear/generalized linear models, where the degrees of freedom are simply the number of parameters.

While PROC GAMPL is a high-performance procedure that executes very fast even in a non-distributed environment, it should be noted that computation speed is highly dependent on the complexity of the model – additional multivariate splines and higher degrees of freedom require extra computing time.

TWO PROPOSED MODELING SOLUTIONS

Now that we have identified the need for an additive model and how best to implement it, we can focus on different methodologies for incorporating additional variables. In this section, we will highlight two approaches for modeling the spatial aspect of our dependent variable.

To conceptualize the difference in these two approaches, we will analyze crime loss data for Homeowners policies in a single state. Using the two methods defined below, we can demonstrate the improvement gained by creating territory rates based on two different implementations of splines as opposed to a locally weighted average. Before we detail these approaches, we must first define our variables and discuss how they are handled in the model.

MODELING SPECIFICATIONS AND DEFINITION OF INSURANCE TERMS

As previously stated, the model target is pure premium (PurePrem), which is defined as loss per exposure in a grid cell. Latitude and longitude combine to create 1km by 1km square grid cells. Other variables included in the modeling process are defined as follows:

- LossExpo – loss exposure. For this Homeowners illustration, we are using earned house years. Values of one mean the policy was in force for the entire term; fractions reflect the portion earned before the policy lapsed.
- Temp – annual temperature range
- Precip – average annual precipitation

To highlight the differences in modeling approaches, the same variables will be used in both methods, and in the same manner (i.e. as linear univariate regressors). The target variable is adjusted to account for all other factors in our rating plan (such as construction type, home security system, etc.). Numerous missing values are present in the target variable, indicating there are no policyholders at the location. While these grid cells are not used to fit the model, they are included in the dataset so that a predicted value is generated. Due to the right-skewed nature of our continuous target, we will use a Gamma distribution with a log link transformation.

METHOD ONE: THE GLM-GAM APPROACH

The GLM-GAM approach is a two stage modeling process. In the first stage, we fit a GLM to our linear terms. In the second stage, we fit a GAM on the same target using the remaining nonlinear spatial variables, and the score of the first model as an offset. Method one is similar in concept to a spatial error model – the main regressors are modeled first, and a second model is run on the residuals to extract the spatial aspect. The advantage of this two stage approach is that it separates the conventional GLM and variable selection process from the spline process, making it easier to implement. The disadvantage is that, in the second model, everything from the first stage model is included in the offset – all the regression coefficients are assumed to be fixed. If the offset has a structure to it from the assumed linear model, and if the original model is over-fitting, this introduces some artificial pattern into the rest of the modeling process.

The HPGENSELECT procedure[4] is utilized for fitting and building our generalized linear model. The SAS code is as follows:

Stage One:

```
PROC HPGENSELECT data = StateX_Crime;
        ID Latitude  Longitude  PurePrem  LossExpo;
        MODEL PurePrem = Temp  Precip  / dist=gamma   link=log;
        WEIGHT LossExpo;
        OUTPUT out=Stage1output   pred=GLM_pred;
RUN;
```

Stage Two:

```
PROC GAMPL data = Stage1output;
        MODEL PurePrem = spline(Latitude Longitude / maxdf=200  maxknots=20000)
                                / dist=gamma   link=log   offset=GLM_pred;
        WEIGHT LossExpo;
        OUTPUT out=Method1output   pred=Method1_pred;
RUN;
```

METHOD TWO: THE SINGLE GAM APPROACH

In contrast to the GLM-GAM approach, a single stage GAM fits all explanatory variables at once, and does not require the non-spatial terms to be linear. The advantage in using a singular approach is that all of the variation is captured. The parameter and smoothing function estimates are calculated jointly rather than fixing the estimates and associated errors at different stages in the modeling process. The disadvantage to this approach is the increased complexity of the model may make it harder to conceptualize and explain. Splines functions consist of multiple basis functions, each with its own parameter, and spline terms are usually interpreted as a single entity (not individual spline basis

parameters). It should be noted that the procedure generates plots of smoothing components as a visual aid for interpreting the fitted spline terms. The SAS code for the Single GAM approach is:

```
PROC GAMPL data = StateX_Crime;
        MODEL PurePrem = spline (Latitude Longitude / maxdf=200 maxknots=20000)
                                 param(Temp)   param(Precip)   / dist=gamma   link=log;
        WEIGHT LossExpo;
        OUTPUT out=Method2output   pred=Method2_pred;
RUN;
```

METHODOLOGY APPLICATION – TERRITORIAL RATEMAKING

The best way to visualize the results is through heat maps, here created using the new HEATMAPPARM statement in the SGPLOT procedure. The city names and points are superimposed using an annotate dataset created from the SAS predefined map datasets[5]. A wider variety of colors indicates more segmentation – an increased ability to achieve a wider range of rates. As is shown in the thermometer to the right of the graphs, dark purple is associated with the lower values, and red with higher.

For the visual comparison, the predicted pure premium values for each grid cell are transformed into relativities. This is done by dividing each grid cell's pure premium by the statewide pure premium, so that a relativity of one is the state average, and values below one denote losses lower that the average for the state.

Our goal is to increase the segmentation in our territorial ratemaking and capture more of the granular variation. We can use heat maps along with the model output to assess the improvement of the two methods over the current approach, as well as to see each methodology's impact on the results.

METHOD COMPARISON

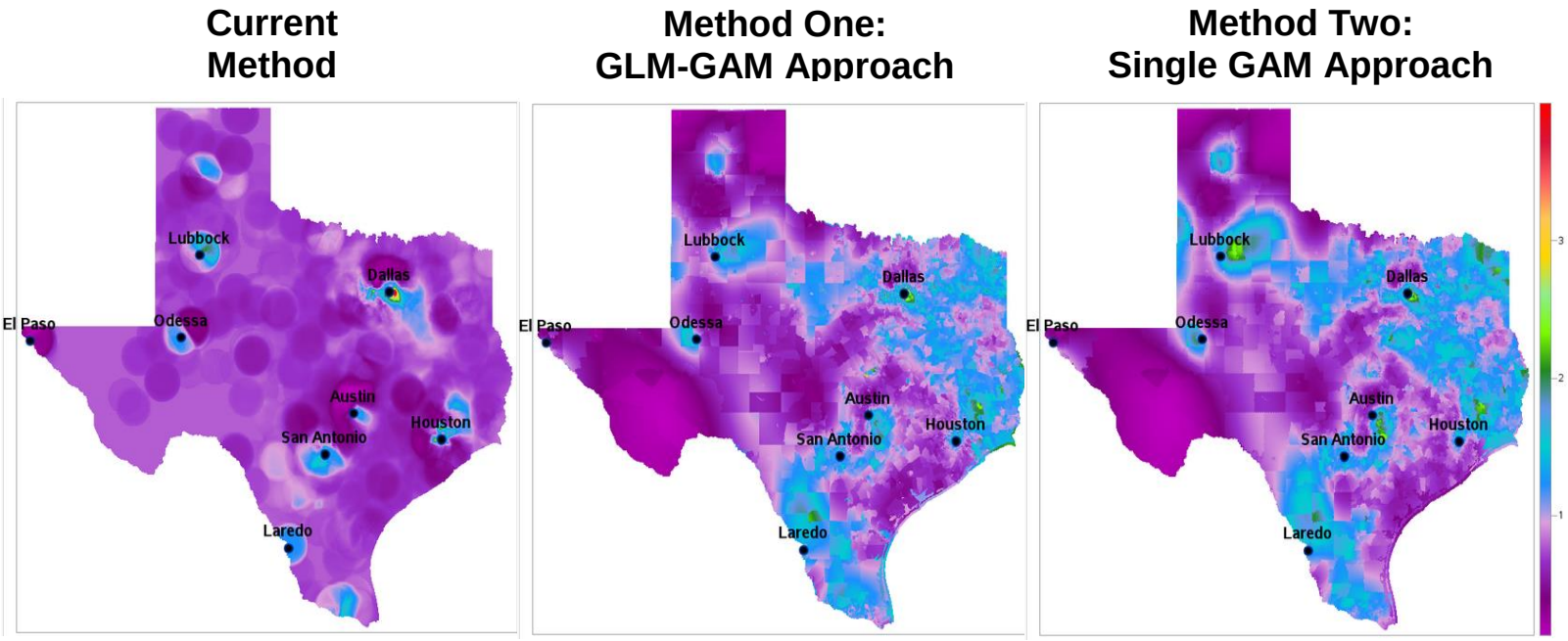
To assess the differences in model fits, we use a few different metrics. GCV (generalized cross validation) is a model fitting criterion used with splines, and is similar to AIC – smaller is better. The metric is relative, lending itself to comparisons of various model runs. The roughness penalty is also a relative metric, measuring the penalization from the functional space bounded by the maximum number of DFs. As in our case, both methods have the same maximum DFs. Therefore, the model with the lower roughness penalty will actually have more parameters (as is more directly seen in the model DF value) since it is penalized less for its complexity.

		Method One	Method Two
Fit Statistics	AIC	204,105	203,791
	GCV	267.92	259.05
	Degrees of Freedom	134.87	144.18
	Roughness Penalty	27,071	26,204
Spline Effects – Latitude & Longitude	Smoothing Parameter	17.82	11.07
	Degrees of Freedom	132.87	140.18
	Roughness Penalty	27,071	26,204

Table 1. Fit Statistics

The plots below demonstrate a dramatic improvement in addressing the granular variation through both the GLM-GAM and Single GAM approaches. Methods one and two still have pockets of higher relativities around the major cities, but they have different degrees of relativity variation for the surrounding areas,

mainly due to the inclusion of additional spatial variables. The difference in the two methods seen in the lower right hand corner of the graph is attributed to using an offset in the GLM-GAM approach – part of the variation has been fixed.



Output 1. Heat Maps

All fit statistics point to the single approach producing the better model. The Single GAM has a lower roughness penalty (lower penalization for its complexity), a smaller GCV criterion (indicating a better fit), and a smaller smoothing parameter (allowing for more granular predictions). The single model is slightly more nonlinear, but more accurately predicts the pure premium, especially along the coast. The Single GAM model results follow the current methodology, with the increased benefit of more granularly modeling the areas between the major cities, as well as within the cities themselves.

CONCLUSION

Generalized additive models offer a flexible way to incorporate smoothing splines for modeling geospatial variables. PROC GAMPL not only aids in the fine tuning of the generalized additive model, but seamlessly allows for the inclusion of parametric terms and additional spline terms – something that is not so effortlessly handled with other implementation approaches. Both the single and two stage approaches produce reasonable results, given the advantages and disadvantages of each methodology. For our particular application, the Single GAM Approach best accomplishes our goal – increased segmentation over the current methodology to better capture the low level variation between policyholders.

REFERENCES

[1]Wood, S. (2006). Generalized Additive Models. Boca Raton, FL: Chapman & Hall/CRC.

[2]SAS Institute Inc. (2015). “The GAMPL Procedure.” In SAS/STAT® 14.1 User’s Guide. Cary, NC: SAS Institute Inc. http://support.sas.com/documentation/onlinedoc/stat/141/hpgam.pdf

[3] SAS Institute Inc. (2015). "The GAMPL Procedure." In SAS/STAT® 14.1 User's Guide. Cary, NC: SAS Institute Inc. http://support.sas.com/documentation/cdl/en/statug/68162/HTML/default/viewer.htm#statug_hpgam_details13.htm

[4]SAS Institute Inc. (2015). "The HPGENSELECT Procedure." In SAS/STAT® 14.1 User's Guide. Cary, NC: SAS Institute Inc. http://support.sas.com/documentation/onlinedoc/stat/141/hpgenselect.pdf

[5] SAS Institute Inc. (2015). In SAS/GRAPH® 9.4: Reference, Fourth Edition. Cary, NC: SAS Institute Inc. http://support.sas.com/documentation/cdl/en/graphref/67881/HTML/default/viewer.htm#p1panwy8lvvc3zn1pjsh20reebaj.htm

ACKNOWLEDGMENTS

Special thank you to Bob Rodriguez and Weijie Cai of the SAS Institute.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the authors at:

Carol Frigo
Property and Casualty Actuarial Department
State Farm Insurance Companies
One State Farm Plaza
Bloomington, Illinois 61710-0001
Carol.Frigo.UJQS@StateFarm.com

Kelsey Osterloo
Property and Casualty Actuarial Department
State Farm Insurance Companies
One State Farm Plaza
Bloomington, Illinois 61710-0001
Kelsey.Osterloo.U15H@StateFarm.com