National to Local to Hyper-Local – Model-Based Localization

Introduction

An increasingly important component of digital and even television and print advertising targeting is location, or more precisely locality. Perhaps the most conspicuous example of such targeting is geo-fencing wherein an advertiser directly pushes ads to mobile devices of persons in the immediate area of the advertisers' outlets. Similarly, cable television and both magazines and newspapers knowing the addresses of their subscribers offer variously precise location-based targeting to their advertising clients. As such, it is increasingly important that the audience measurement services supporting such media offer their audience and targeting measures at relatively granular geographic levels.

This need for increasingly granular local information poses a genuine challenge to media audience measurement services chiefly due to issues related to sample (or as in the case of cable television return-path data) coverage. For example, in the U.S. an adequate sample for a Designated Market Ares [DMA] or even a county within the DMA may have little or no coverage for individual postal codes or neighborhoods, etc. Moreover the costs of increasing such samples to provide such coverage are often prohibitive.

Beyond local media audience measurement services, for many decades geographic-based demographic segmentation systems have offered local and hyper-local metrics. Similarly, at the DMA level GfK MRI a number of years ago introduced a borrowed strength based dataset offering local area estimates for its generally nationally (and regionally and top market) based measures.

The focus of the work at hand is to develop a more dynamic, model-based approach to small area estimation of measures gathered from what is essentially a nationally designed sample – GfK MRI's Survey of the American Consumer [SAC]. In particular, this work employs geographically weighted regression [GWR] wherein the resulting regression model is sensitive to locality, in the case at hand, defined by latitude and longitude – i.e. accounts for "where locally weighted regression coefficients move away from their global values." [1] ["Geographically Weighted Regression", Roger Bivand, 2017]

In broad brushstrokes GWR's will be developed for a number of publications from SAC and evaluated against comparable non-geographic multivariate regressions using cross-validation for hold-out samples.

Fully realized, this strategy for local and hyper-local media and consumer target estimation possesses the accuracy of specific model-based estimation, the precision of accounting for locally dependent variability, the robustness of large sample sizes and the convenience of "just-in-time" estimation.

Geographically Weighted Regression Models – An Overview

Geographically weighted models [GWR] are useful in contexts where data with some spatial orientation are not well characterized by a global model, "but where there are spatial regions where a suitably localized model calibration provides a better description" [2]. Broadly, the approach utilized by GWR is to develop a series of localized models defined by a spatially determined window whose size is characterized by a bandwidth embodying a decay function which weights the cases appropriately [3], i.e. with respect to their proximity to the centroid of windowed area. In fact, a primary aspect of the work of the GWR is the development of a suitable bandwidth avoiding the twin perils of highly localized but noisy data on the one hand and over generality on the other.

In the work at hand, the bandwidths of the various models are developed by optimizing a Gaussian function wherein relative to a series of calibration points cases are weighting with decreasing weights with respect to their distance from the calibration point. [4] The optimization seeks to minimize the cross-validation error as determined by the root mean square error of the prediction.

GfK MRI Data

GfK MRI's SAC employs an area probability sample design with stratification by household income and county size (for additional information consult: http://www.mri.gfk.com/gfk-mri-technical-guides). In weighting SAC, first a design weight is developed to account for the differential probability of selection and then a sample balance is performed to develop the final population projection weights.

For the GWR testing GfK MRI's Spring 2015 SAC was employed with a respondent count of 23,978. The SAC data was divided into Training and Test groups each of 11989 cases respectively. The selection of the Training and Test groups was done at random but with probabilities of selection which were approximately inversely proportional to sample incidence at the state/county (FIPS) level. While not completely precise this attempts to develop a training set

with fairly even geographic dispersion, i.e. less subject to the stratification and resultant concentration of the basic SAC sample design.

For the purposes of this initial exploration a series of GWR models were generated for the reading (probability of reading developed from the read/screen ratio at each frequency level) of ten publications. While not exhaustive with respect to print genre, the ten publications included those with particular demographic (e.g. Gender, Age Ethnicity), geographic (e.g. regional publications), interest (e.g. sports, women's service) distributions along with general editorial and a national newspaper.

In developing the GWR's a set of ten (10) demographic and geographic independent variables were employed – Gender, Age, African-American/Yes-No, Hispanic/Yes-No, Household Size, Household Income, Education, Presence of Children in Household Census Division and County Size. With respect to the latter two geographic measures, their inclusion serves to isolate the incremental value of the GWR approach.

While the methods herein explored can apply to any media or non-media target, the context for this particular exercise encourages the application to media and particularly print media.

Geographically Weighted Regression Analysis Software

Much of the work of this analysis employs the R programming language and in particular the spgwr package [5]. (Note, other R packages, e.g. GWModel, are available to undertake GWR work and may be evaluated at a later date as part of a more extended effort.) Two spgwr package functions in particular were extensively employed. The first, gwr.sel was used to find the optimal bandwidth (Gauss – gweight=gwr.Gauss) for the linear models. Once the optimal bandwidth had been determined gwr was employed to execute the GWR with the optimized bandwidth and generate the model.

Geographically Weighted Regression – Optimal Bandwidth

As noted, the first step in GWR is the determination of a bandwidth for the localized models. The GfK MRI SAC dataset includes the latitude and longitude for the FIPS (state/'county code) for all respondents. (Greater specificity of geographic coordinates may have been productive but was unavailable for this dataset.) The bandwidth selection optimization essentially performs an extensive series of regressions using the Training dataset with the reading (probability of reading) of each publication as the dependent variable, the independent variables as noted previously and with successively more optimal geographic bandwidths determining the case weighting.

Considering the ten publications noted previously the optimized bandwidth (i.e. geographic expanse) ranged from 3.30 to 58.35 kilometers with a mean of 16.43.

Table #1 – Optimal Bandwidths

Publication	Optimal Bandwidth
General Editorial #1 – Older Age Skew	6.29
General Editorial #2	7.27
Women's Service	8.71
Epicurean	58.35
General Editorial #3 – Ethnic Skew	24.00
General Editorial #4 - Regional Skew	3.30
Entertainment/Celebrity	46.51
Spanish Language	3.30
General Editorial – #5 Regional Skew	3.30
National Newspaper – Sunday Edition	3.30

A few general comments about the differences in the range of bandwidths are worthwhile:

1) The larger bandwidths are associated with publications for which several of the independent variables carry substantial explanatory force, where one of those variables is related to geography (e.g. Census Division and/or County Size) and where there is minimal variance among the range of geographically dependent estimates for that geographic variable. For example, the Epicurean publication has a strong Female skew, the Southern United States Census Divisions are significant and consistent across the bandwidth dependent regressions. In short, highly localized regressions (i.e. ones with minimal bandwidths) contribute little beyond what is explained by Census Division and County Size.

2) Smaller bandwidths result when the independent variables are weaker with respect to accounting for localized variation. For example, for the Spanish Language periodical, Hispanic is, as expected, a strong general predictor, but neither Census Division nor County Size are consistent predictors. In short, there is substantial local variation unexplained by these two more general geographic independent variables.

Geographically Weighted Regression – Prediction

Once the optimal bandwidth is determined GWR proceeds by iteratively executing a series of linear regressions but with cases weighted by proximity to successive bandwidth windows. In short, GWR is really a series of linear regressions performed on geographically proximate cases in this cases from the Training dataset. The result is that rather than a single set of (global) regression coefficients each case of the Training dataset has its own set of coefficients resulting from the aggregation of the coefficients developed from the successive bandwidth controlled regressions.

Table #2 – GWR Regression Coefficients

	Min.	1st Qu.	Median	3rd Qu.	Max.	Global
X.Intercept.	-0.18130433	-0.11447807	-0.07608781	-0.03384880	0.02776294	-0.0773
PDemoAge	0.00471500	0.00626347	0.00722107	0.00774890	0.00836692	0.0070
PDemoSexM	-0.04749964	-0.03667577	-0.03057717	-0.02776202	-0.02485156	-0.0322
PDemoHispanicNon	-0.03378954	-0.02043369	-0.01814266	-0.01509563	0.00319724	-0.0172
PDemoRaceOther	-0.10343374	-0.09505580	-0.09150589	-0.08711674	-0.04850373	-0.0883
GDemoCensusDivESC	-0.05860407	-0.02087529	-0.00654385	-0.00117760	0.01765644	-0.0059
GDemoCensusDivMA	-0.03617161	-0.00601867	0.01393326	0.02531928	0.03508604	0.0224
GDemoCensusDivMNT	-0.03575365	0.01296578	0.02615679	0.02937024	0.03389928	0.0193
GDemoCensusDivNE	-0.02956131	-0.00143935	0.01448933	0.01967686	0.02375587	0.0170
GDemoCensusDivPAC	-0.07354759	-0.02821875	-0.01968197	-0.01261358	-0.00472305	-0.0261
GDemoCensusDivSA	-0.05121228	-0.02242156	-0.00520010	0.01364681	0.02849032	-0.0006
GDemoCensusDivWNC	-0.09537575	-0.02886466	0.00229248	0.01492428	0.02264879	-0.0076
GDemoCensusDivWSC	-0.09877657	-0.03967254	-0.02155663	-0.00923836	-0.00179788	-0.0298
HDemoHHSize	-0.00707073	0.00069322	0.00646688	0.01329362	0.02288706	0.0068
HDemoHHI	-0.00546269	-0.00113554	0.00025585	0.00053773	0.00192128	-0.0004
PDemoEduc	-0.00207838	0.00645924	0.00790179	0.00900671	0.01503073	0.0071
HDemoChildYes	-0.09276648	-0.07223773	-0.06118824	-0.05624621	-0.05176126	-0.0656
HDemoCntySize	-0.01801263	-0.01324924	-0.00952801	-0.00159138	0.02807420	-0.0055

Table #2 shows the summary of the range of coefficient values (and intercept) for a GWR of one of the publications along with the coefficient for a standard linear regression ("Global"). The variance for each of the GWR independent variables (and the difference from the global coefficients) suggests the extent to which geographic variance persists beyond that accounted for by the Census Division and County Size geographic variables.

Finally, a set of predictions for the SAC Test dataset was generated. To develop these predictions each case of the Test dataset was scored using the mean of the coefficients for the cases in the Train dataset who were within the GWR's bandwidth. In cases where there were fewer than ten (10) such Training cases the bandwidth was increased until at least ten Training cases were available. The predicted values were then compared through correlation with the actual reading probability values.

Table #3 – GWR versus Standard Linear Regression

Publication	GWR	Standard/Global	GWR/Standard
General Editorial #1 – Older Age Skew	0.4016	0.3984	0.9909
General Editorial #2	0.1110	0.1083	0.9218
Women's Service	0.3556	0.3541	0.9939
Epicurean	0.1450	0.1448	0.9999
General Editorial #3 – Ethnic Skew	0.5149	0.5148	0.9997
General Editorial #4 - Regional Skew	0.2184	0.1927	0.8585
Entertainment/Celebrity	0.3123	0.3122	0.9999
Spanish Language	0.2665	0.2495	0.8954
General Editorial – #5 Regional Skew	0.3741	0.3276	0.9032
National Newspaper – Sunday Edition	0.2587	0.2131	0.8361

As is evident from Table #3, the results from the GWR were (very) modestly, but consistently better than comparable results using the standard linear regression (Global). Of particular import, the two publications with decidedly regional skews and the national newspaper, also having a strong regional skew, expectedly benefit the most from GWR.

Conclusions and Future Work

As noted in the introduction localized and/or location-based metrics for media and consumer behaviors are increasingly important. While some highly localized measures are available they are usually incomplete with respect to critical behavioral (and even demographic) measures. Conversely, audience and consumer measurement products such as GfK MRI's SAC often do possess these critical behavioral (and demographic) measures but are not conventionally projectable to small geographic areas. Hence, the work at hand is best understood as an initial attempt to identify a path from the more geographically general to the more localized using relatively dynamic methods.

Endnotes

1. Bivand, Roger, "Geographically Weighted Regression", (2017).
2. Gollini I, Lu B, Charlton M, Brunsdon C, Harris P (2015). "GWModel: an R Package for Exploring Spatial Heterogeneity using Geographically Weighted Models." Journal of Statistical Software, 63 (17), 1.
3. Ibid, page 1
4. Ibid, page 5.
5. Roger Bivand and Danlin Yu (2008). spgwr: Geographically weighted regression. R package version 0.6-31.

Authors

James Collins is Executive Vice President, Research at GfK MRI where he focuses on data integration. A frequent contributor to PDRF and other conferences, Collins' primary research foci beyond data integration are optimization and systems development.