Bayesian Analysis (2016) **11**, Number 1, pp. 299–302

Comment on Article by Page and Quintana*

Robert B. Gramacy† and Herbert K. H. Lee‡

We congratulate Page and Quintana for this new approach to nonstationary spatial modeling. This new model is based on a spatial clustering, whereby the underlying clusters are defined using spatial structure, and these clusters help drive the correlations between the observations. Here we discuss several related models that are derived from regional partitions or neighborhood structures. We also note that this work falls under the much larger rubric of nonstationary spatial modeling, and there are quite a number of other approaches that are related in that context.

1 Regional Partition Models

A related approach in the literature is to define regional partitions, rather than potentially overlapping clusters. A proper partitioning can be computationally simpler, in that the combinatoric possibilities are substantially more limited than for a clustering approach. The advantage of a clustering approach is additional flexibility in defining the clusters, including the allowance of irregular shapes and overlapping clusters. Thus the key trade-off is whether the additional computational cost is justified by the need for modeling flexibility.

Given the opening example in Figure 1 of the paper, of spatial fields that change across rectangular regions, we were surprised that there is no mention of treed Gaussian process (TGP) models (Gramacy and Lee, 2008), as they fit this example exactly. A TGP model considers a tree-based partitioning of the space and fits independent Gaussian process models within each partition. By allowing the partition structure to also be a random variable inferred simultaneously, Bayesian model averaging can result in smooth predictions when the data are smooth (as is the typical case), yet can provide for sharp jumps if warranted by the data (such as in Figure 1). With the inherent flexibility of the Gaussian process, just a little nonstationarity is usually enough to provide a really good fit to data, and this partitioning approach provides that sufficient amount of nonstationarity without invoking the massive computational burdens of fully nonstationary models. Open source software is provided in the `tgp` package (Gramacy, 2007; Gramacy and Taddy, 2010) for R (R Core Team, 2015).

Other examples of regional partition-based models are tessellations and partitioned process convolutions. Kim et al. (2005) developed a model that partitioned the space using a tessellation, and then fit independent Gaussian processes within each of those partitions. This provides additional flexibility beyond the treed partitioning approach,

*Main article DOI: 10.1214/15-BA971.
†Booth School of Business, University of Chicago, Chicago, IL, rbgramacy@chicagobooth.edu
‡Department of Applied Mathematics and Statistics, Baskin School of Engineering, University of California, Santa Cruz, CA, herbie@soe.ucsc.edu

 DOI: 10.1214/15-BA971A

but in practice this additional flexibility does not usually much improve the fit or prediction, because the Gaussian process is already quite flexible. A different partitioning mechanism (Liang and Lee, 2011) is to start with the process convolution representation of a Gaussian process, wherein a smoothing kernel is convolved with a white noise background process to create a Gaussian process, and then allow partitioning for the kernel parameters and the background processes. This approach guarantees that the model will produce a continuous response surface (assuming a smooth kernel), but allows regional variability in model structure, again providing nonstationarity computationally cheaply. The process convolution representation allows the fitting of much larger datasets without the standard explosion in computational expense, but is somewhat limited to lower-dimensional input spaces.

2 Neighborhood Models

Another, perhaps more fluid, localized inference approach involves approximating the predictive equations nearby elements of a predictive (or testing) input set. Cressie called this "ad hoc" *local kriging* (e.g., Cressie, 1991, pp. 131–134), however the idea has come a long way since then. Emory (2009) is the most recent author to have considered the problem in a spatial statistics context, while Gramacy and Apley (2015) focus primarily on computer experiments, and provide open source software in the form of an R package called `laGP` (Gramacy, 2015). The above works are primarily motivated by addressing computational issues that are faced with dealing with large data sets, specifically "large N" problems for training data sets with N records. Gaussian process inference and prediction require $O(N^3)$ dense matrix decompositions in that case, which is computationally intactible when $N \gg 1000$. Instead, they suggest that a subset of the data $X_n(x)$ of size $n \ll N$, nearby to predictive locations x could lead to nearly identical predictions, compared to the full data set, with a fraction of the computational effort. It turns out that a nearest-neighbor choice of $X_n(x)$, as advocated e.g., by Emory (2009) is inefficient. Gramacy and Apley (2015) show that a sequential search for locations which minimize mean-squared prediction error works better, yet has similar computational demands. They also recognize the nonstationary modeling potential of local inference, and demonstrate more accurate predictions compared to a more spatially homogeneous (neighborhood-based) approach. Both are examples of *local approximate* Gaussian process (laGP) predictors, and one can envisage many further variations.

Another way to build local neighborhoods is through the use of covariance tapering, for example in Anderes et al. (2013) where tapering allows for efficient fitting of local variation in the correlation structure. Similar to how Gramacy and Apley (2015) enhance the fidelity of local neighborhood models to make them more nonstationary, Anderes et al. (2013) can be interpreted as doing the same for a similar more stationary approach that was primarily focused on remedies for big data problems (Kaufman et al., 2012).

3 Chilean Standardized Testing Data

The authors graciously provided access to the Chilean Standardized Testing data so that we could run a comparison of TGP and laGP, and ordinary stationary Gaussian

process (GP) models, with spatial product partition models. We follow the setup of Section 4.2 for the Conditional Model, fitting unstandardized SIMCE on spatial location and mother's education scores. Using the first 600 datapoints as training data and the remaining 615 observations as test data, both TGP and an ordinary stationary GP provide a mean square prediction error (MSPE) of about 365, which put them ahead of everything but CPS C_4. We note that this error rate is highly dependent on the particular observations used for the training and test dataset. We also tried taking 100 random splits between training and test datasets and obtained MSPEs for a stationary GP ranging from 304.6 to 403.6, with an average of 349.9.

Looking more closely at the fitted models for both TGP and laGP, as well as a stationary GP, turned up a surprising result. Our TGP model did not partition the space, but was using a stationary Gaussian process across all of the observations. laGP grew its local neighborhood so that it wasn't fitting a particularly local model. Thus in both cases, these flexible models reverted toward fitting a stationary GP, or nearly so. Moreover, we entertained a 10-fold cross-validation comparing laGP and stationary GPs in 100 repetitions. Although the resulting MSPEs were very similar for both approaches, having nearly the same marginal mean and variance across all 1000 testing sets, a pairwise t-test resoundingly rejected the null that the two approaches were yielding the "same" results. Indeed, the stationary GP was consistently yielding lower MSPEs, albeit not by a large margin. So we are compelled to ask the authors how much benefit SPPM provides over a stationary GP on this dataset.

4 A Final Thought

Since clustering can often have useful interpretations in the input space, have the authors given any thought to trying to interpret the clusters that result from their model? That could be useful for better understanding the structure of a dataset.

References

Anderes, E., Huser, R., Nychka, D. W., and Coram, M. (2013). "Nonstationary Positive Definite Tapering on the Plane." *Journal of Computational and Graphical Statistics*, 22(4): 848–865. MR3173746. doi: http://dx.doi.org/10.1080/10618600.2012.729982. 300

Cressie, N. (1991). *Statistics for Spatial Data, revised edition*. John Wiley and Sons, Inc. MR1127423. 300

Emory, X. (2009). "The kriging update equations and their application to the selection of neighboring data." *Computational Geosciences*, 13(3): 269–280. 300

Gramacy, R. (2015). "`laGP`: Large-Scale Spatial Modeling via Local Approximate Gaussian Processes in R." Technical report, The University of Chicago. Available as a vignette in the `laGP` package. 300

Gramacy, R. B. (2007). "`tgp`: An R Package for Bayesian Nonstationary, Semiparametric

Nonlinear Regression and Design by Treed Gaussian Process Models." *Journal of Statistical Software*, 19(9): 1–46. http://www.jstatsoft.org/v19/i09/. 299

Gramacy, R. B. and Apley, D. W. (2015). "Local Gaussian process approximation for large computer experiments." *Journal of Computational and Graphical Statistics*, 24(2): 561–578. MR3357395. doi: http://dx.doi.org/10.1080/10618600.2014.914442. 300

Gramacy, R. B. and Lee, H. K. H. (2008). "Bayesian Treed Gaussian Process Models with an Application to Computer Modeling." *Journal of the American Statistical Association*, 103: 1119–1130. MR2528830. doi: http://dx.doi.org/10.1198/016214508000000689. 299

Gramacy, R. B. and Taddy, M. (2010). "Categorical Inputs, Sensitivity Analysis, Optimization and Importance Tempering with tgp Version 2, an R Package for Treed Gaussian Process Models." *Journal of Statistical Software*, 33(6): 1–48. http://www.jstatsoft.org/v33/i06/ MR2578072. doi: http://dx.doi.org/10.1007/s11222-008-9108-5. 299

Kaufman, C., Bingham, D., Habib, S., Heitmann, K., and Frieman, J. (2012). "Efficient Emulators of Computer Experiments Using Compactly Supported Correlation Functions, With An Application to Cosmology." *Annals of Applied Statistics*, 5(4): 2470–2492. MR2907123. doi: http://dx.doi.org/10.1214/11-AOAS489. 300

Kim, H.-M., Mallick, B. K., and Holmes, C. C. (2005). "Analyzing Nonstationary Spatial Data Using Piecewise Gaussian Processes." *Journal of the American Statistical Association*, 100: 653–668. MR2160567. doi: http://dx.doi.org/10.1198/016214504000002014. 299

Liang, W. W. J. and Lee, H. K. H. (2011). "Bayesian Nonstationary Gaussian Process Models For Large Datasets Via Treed Process Convolutions." Technical Report UCSC-SOE-11-25, University of California, Santa Cruz, School of Engineering. 300

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/. 299