Ecological Archives **E091-048-A2**

Erin E. Peterson and Jay M. Ver Hoef. 2010. A mixed-model moving-average approach to geostatistical modeling in stream networks. *Ecology* 91:644–651.

Appendix B: Additional details for the analysis of the PONSE data set.

South East Queensland (SEQ) is located on the eastern coast of Australia and is approximately 22,999 square km in size. It is a subtropical region with mean annual maximum temperatures ranging between 21 and 29 °C (EHMP, 2001) and total annual rainfall ranges between 900 and 1800 mm (Pusey et al. 1993). Elevation ranges between 0 meters in coastal areas to 1360 meters in the west along the Great Dividing Range. Regional discharge is seasonally variable and is related to elevation, slope, and rainfall (EHMP, 2001). The predominant land uses in the region are natural bushland (37%) and grazing (35%).

The Ecosystem Health Monitoring Program (EHMP) has been collecting indicators of biotic structure and ecosystem function throughout SEQ since 2002 (Bunn et al. *in press*). The program aims are to evaluate the condition and trend in ecological health of freshwater environments and to guide investments in watershed protection and rehabilitation. Metrics based on fish assemblage structure and function are commonly used as indicators of ecological health because they are thought to provide a holistic approach to assessment across broad spatial and temporal scales (Harris 1995). The EHMP uses ecosystem health indicators based on freshwater fish species richness (Kennard et al. 2006*a*), fish assemblage composition (Kennard et al. 2006*b*) and the relative abundance of alien species (Kennard et al. 2005). In this example, we used a model-based fish indicator, the proportion of native fish species expected (PONSE), which is the ratio of observed to expected native fish species richness (Bunn et al. *in press*, Kennard et al. 2006*a*). Here, species richness is simply the number of native fish species observed or predicted

at a location. Native species richness is a commonly used summary indicator of ecosystem health that may reflect a variety of disturbances functioning at a range of spatial and temporal scales. Environmental degradation is expected to alter naturally diverse fish communities to simple assemblages dominated by only a few tolerant species. Therefore, species richness is expected to decline with increasing environmental stress. For example, Kennard and others (2006*a*) showed that native species richness was lower than expected at test sites in SEQ affected by intensive watershed land use (clearing, grazing, cropping and urbanization), degraded water quality (high diel temperature range, low pH, high conductivity, high turbidity) and riparian and aquatic habitat degradation.

The observed species richness values used to calculate the PONSE scores were based on fish assemblage data collected at 86 survey sites during the spring of 2005; hereafter referred to as the "observed sites". Fish assemblages were sampled using a combination of backpack electrofishing and seine-netting where possible (Kennard et al. 2006*c*). Electrofishing was conducted using a Smith-Root model 12B Backpack Electroshocker, while seine-netting was conducted using a 10m long (1.5m drop) pocket seine of 10mm (stretched) mesh. An attempt was made to intensively sample all habitat unit types (pool, riffle, and run) at each site (Kennard et al. 2006*c*). When only one habitat unit type was present, at least two units were fished. The average length of fished stream was 75m and data from the entire length were combined to obtain an observed species richness value.

The expected species richness values used to calculate each PONSE score were generated using a referential model and represent the number of native species that are expected to be present in a physically similar, but undisturbed stream (Kennard et al. 2006*a*). A regression tree was used to partition a large set of minimally-disturbed reference sites into homogenous groups

based on environmental characteristics including elevation, distance to the stream outlet, distance from the stream source, and the mean wetted stream width (Kennard et al. 2006*a*). The mean species richness for each reference site group was then calculated. The 86 observed sites were assigned to reference site groups based on the same environmental characteristics mentioned previously. Finally, each observed site was assigned an expected species richness value, which was equal to the mean species richness for the reference site group. For additional details about the calculation of the expected species richness values, please see Kennard and others (2006*a*) and EHMP (2001).

We generated the spatial data necessary for geostatistical modeling in a geographic information system (GIS) using ArcGIS version 9.2 software (ESRI 2006). The Functional Linkage of Waterbasins and Streams (FLoWS) toolset (Theobald et al. 2006) was used to construct a landscape network, which is a spatial data structure designed to store topological relationships between nodes, directed edges, and polygons (Theobald et al. 2005). Here, we used the landscape network to represent stream segments (directed edges) and confluences (nodes). The location of additional features, such as survey sites, can also be associated with the data structure. We incorporated the 86 observed sites and 137 non-randomly selected "prediction sites" (where PONSE scores were not measured) into the landscape network. Scripts written in Python version 2.4.1 (Van Rossum and Drake 2003) were used to generate the hydrologic distances and spatial weights between all observed and prediction sites based on the topological relationships stored in the landscape network. The spatial weights were based on watershed area, which was used as a surrogate for discharge. This seemed like a viable alternative since mean annual discharge has been shown to be correlated with watershed area in other regions (Vogel et al. 1999).

Watershed-scale explanatory variables for each observed and prediction site were calculated using a combination of FLoWS tools (Theobald et al. 2006) and customized scripts written in Python version 2.4.1 (Van Rossum and Drake 2003). The streams data were provided by the Moreton Bay Waterways and Catchments Partnership (2005); here we considered a single GIS polyline feature as a stream segment. A digital elevation model (Queensland Natural Resources and Water 2000) and the Queensland Land Use Mapping Program data set (Bureau of Rural Sciences 2002) were used to generate seven watershed explanatory variables (Table B1). The EHMP classified streams based on elevation, mean annual rainfall, stream order, and stream gradient (EHMP 2001) to create four EHMP regions: Tannin-stained, Coastal, Lowland, and Upland, which we also included as a site-scale explanatory variable (Table B1). We checked the explanatory variables for collinearity using the variance inflation factor statistic (Helsel and Hirsch 1992), which indicated that it was unnecessary to remove explanatory variables. Note that all of the statistical analyses described here were performed in R version 2.6.1 (R Development Core Team 2008).

We used a two-step model selection procedure to compare models. First we fixed the covariance structure and focused on the explanatory variables. All of the mixed models were fit using the exponential tail-up (TU), linear-with-sill tail-down (TD), and Gaussian Euclidean (EUC) as component models (formulas for component models are given in Appendix A). The choice of autocovariance function for the tail-up and Euclidean components was somewhat arbitrary because only one type of spatial relationship is represented by each of the functions (flow-connected or Euclidean). Differences in the strength of spatial autocorrelation between locations when different autocovariance functions are fit to the data result from the shape of the autocovariance function and are likely to be minimal. In the case of the tail-down model the

difference in performance between autocovariance functions is attributed to both the shape of the function *and* the way in which flow-connected and flow-unconnected relationships are represented. More specifically, the relative strength of spatial autocorrelation for each relationship differs depending on the autocovariance function that is used. For example, consider a case where there are two pairs of locations. One pair is flow-unconnected and the other is flow-connected and the distance between pairs is equal ($a+b=h$). If the tail-down exponential model is fit to the data the strength of spatial autocorrelation between the two pairs is equal. However, if the linear-with-sill function is fit to the data the strength of spatial autocorrelation between the flow-unconnected pair may be up to 1.5 times that of the flow-connected pair. We chose the linear-with-sill tail-down autocovariance function because we wanted to model maximum autocorrelation among flow-unconnected sites relative to flow-connected sites. All tail-up models have zero autocorrelation among flow-unconnected sites, so any function allows us to model minimum autocorrelation among flow-unconnected sites relative to flow-connected sites. The combination of linear-with-sill tail-down plus any tail-up model in a mixed model allows the most flexible modeling of autocorrelation. More details may be found in Ver Hoef and Peterson (*in press*).

We estimated all of the parameters in the first phase of model selection using maximum likelihood. Outliers can be overly influential when likelihoods are calculated (Martin 1980) and when the covariance function is fit to the data (Cressie 1993). Therefore, a backwards stepwise model selection strategy was implemented so that the model diagnostics could be examined for every model. Nevertheless, no outliers were identified or removed during the analysis. We selected explanatory variables based on the smallest Akaike Information Criterion (AIC) (Akaike 1974) for the fitted models to determine which had the most support in the data.

During the second phase of model selection we focused on selecting the most appropriate covariance structure. We used the selected explanatory variable set as described above to compare every linear combination of TU, TD, and EUC covariance structure, where four different autocovariance functions were tested for each model type. For the TU and TD models, these included the spherical, exponential, mariah, and linear-with-sill functions (formulas are given in Appendix A). The EUC model included the spherical, exponential, Gaussian, and Cauchy functions (Chiles and Delfiner 1999). This resulted in a total of 124 models, each with a different covariance structure. In addition, we fit a classical linear model assuming independence to compare to models that use spatial autocorrelation. Maximum likelihood may produce biased estimates of the covariance parameters and so we used restricted maximum likelihood (REML) for parameter estimation (Cressie 1993). Note that REML was not used for parameter estimation in the first phase of model selection because a side effect of REML is that information criteria, such as AIC, can only be used if the explanatory variable set remains fixed (Verbeke and Molenberghs 2000). Once the fitted covariance matrix had been generated using REML, it was then used to estimate the fixed effects. This is referred to as "empirical" best linear unbiased estimation and is often used in software such as SAS (Littell et al. 1996).

Leave-one-out cross-validation predictions were generated at the observed sites for each of the 124 models using universal kriging (Cressie 1993) and then used to calculate the root-mean-squared-prediction error (RMSPE),

$$\text{RMSPE} = \sqrt{\sum_{i=1}^{n} (\hat{z}_i - z_i)^2 / n} \qquad \text{(B.1)}$$

where $\hat{z}_i$ is the prediction of the *i*th datum after removing it from the observed data set. The RMSPE can be thought of as proportional to the prediction interval and its interpretation is fairly

straight-forward. For example, if the RMPSE for the best model is one third that of a competing model the gain in using the best model is 66%. The RMSPE was used to compare the models based on different covariance structures. Other model selection criteria, such as information theoretic measures (Burnham and Anderson 1998), could also have been used. The method that we chose met our purpose, which was to identify the covariance mixture that provided the best predictions and it is not our intention to debate the merits of various approaches here. Once the model with the smallest RMSPE was identified, universal kriging was used to make predictions at the 137 prediction sites.

An empirical semivariogram of the model residuals, divided into flow-connected and flow-unconnected relationships (Fig. B1), was generated using the classical estimator as given in Cressie (1993). We also used the covariance parameter estimates to calculate the percent of the variance explained by each of the covariance components (%VC)

$$\%VC_{TU} = \frac{\sigma^2_{TU}}{\sigma^2_{TD} + \sigma^2_{TU} + \sigma^2_{NUG}} * 100, \qquad \text{(B.2)}$$

where σ^2_{TU} and σ^2_{TD} are the partial sill parameters for the TU and TD models and σ^2_{NUG} is the nugget effect. Eq. B.2 demonstrates how to calculate the %VC for the TU model, which in this example is part of a two-component mixture model. However, the %VC can be calculated for any number of model components and the nugget effect. Since it is a percentage, the %VC for all of the components should always sum to 100.

The lowest RMSPE value was produced by the exponential TU/linear-with-sill TD mixture model, which we will hereafter refer to as the final model. The fixed effects for the final model are provided in Table B2. Note that the parameter estimates in Table B2 were generated using REML during the second phase of model selection, but the specific fixed effects were chosen in the first phase of model selection. The final model only contained one explanatory variable, mean slope in the watershed, which was positively correlated with PONSE. This

statistical relationship may represent a physical relationship between PONSE and an anthropogenic disturbance gradient related to land use, water quality, channel or riparian condition, or in-stream habitat (Kennard et al. 2006*a*), rather than slope itself. For example, watersheds with steeper slopes might be expected to have less cleared or cropped land and, as a result higher PONSE scores. In addition, steep slopes are more likely to be found in headwater streams, which may be inaccessible to alien species if they are upstream of barriers to fish movement (Kennard et al. 2005). However, we did not have access to extensive information about anthropogenic disturbances at EHMP sites and were unable to specifically account for these effects in our model.

An empirical semivariogram of the final model residuals, divided into flow-unconnected and flow-connected relationships is shown in Fig. B1. The partial sill and range parameters for the TU component were 0.0162 and 1160.56 km, respectively. In contrast, the TD model had a larger partial sill (0.0464) and a smaller range parameter (110.67 km) than the TU model. The nugget effect was 0.0096. The TU range parameter was substantially larger than the largest flow-connected separation distance (262.19 km). The magnitude of the TU range parameter simply indicates that flow-connected sites are spatially correlated regardless of their location in the stream network. Covariance parameters are sometimes visually estimated from the semivariogram, but it would not be appropriate in this case since there is no weighting for the flow-connected locations. The weights affect the autocorrelation between sites and so there is no guarantee that sites close in space have a strong influence on one another if there is a confluence between them. Calculating the percent of the covariance explained by each of the components provides additional information about the covariance structure of the models. In the final model the TU, TD, and NUG components explained 22.43%, 64.32%, and 13.25% of the variance, respectively.

Our results show that a model based on a mixture of covariances produced the most precise PONSE predictions (Fig. B2). Models based on a covariance mixture tended to produce smaller RMSPE values than models based on a single covariance type. When more than one covariance

type was incorporated, mixture models that included the TD model outperformed other mixture types. There appeared to be more within model type variation in the TD model compared to the TU and EUC model types (Fig. B2). As we discussed previously, this reflects differences in model fit related to both the shape of the autocovariance function and the way that spatial autocorrelation is represented between flow-connected and flow-unconnected pairs. All of the spatial models out performed the classical linear model, with the final model demonstrating an 18.86% gain based on the RMSPE.

Given that the final model only demonstrated a 1.17% gain in predictive ability on the TD model, some might question the usefulness of using a mixture model. Nevertheless, it is unclear which covariance type or mixture will provide the best model fit in advance. A geostatistical methodology is used to model the spatial structure in the residual error and so the observable patterns may not reflect the pattern or scale that would be expected. For example, the unexplained variability in fish distribution may be related to a strongly influential explanatory variable, such as elevation, that was not proposed during model selection. In that case, Euclidean patterns of spatial autocorrelation in fish distribution (resulting from elevation) might be observed at broad spatial scales even though fish movement is generally restricted to the stream network and some species may not migrate large distances. This is a simple example of how environmental factors can produce patterns of spatial autocorrelation; in a freshwater environment, extremely complex process interactions occurring within and between the stream and the terrestrial environment would be expected. Therefore, we recommend using a full covariance mixture (TU/TD/EUC) rather than attempting to make an a priori assumption about the covariance structure of the data. Although the full covariance mixture was not the best model in this example, the loss in predictive ability was only 0.34% when it was used. These results demonstrate an important feature of the covariance mixture approach; namely, that it is flexible enough to represent the entire range of covariance structures (i.e., single EUC, TU, or TD or any combination of the three). This flexibility also makes the approach suitable for a wide variety of data sets collected within a stream network.

LITERATURE CITED

Akaike, H. 1974. A new look at the statistical model identification. *IEEE Transactions on Automatic Control* 19 (6):716–722.

BRS (Bureau of Rural Sciences). 2002. Land use Mapping at Catchment Scale: Principles, Procedures and Definitions, Edition 2. Bureau of Rural Sciences, Department of Agriculture, Fisheries and Forestry – Australia. Kingston, ACT, Australia. p 46.

Bunn, S., E. Abal, M. Smith, S. Choy, C. Fellows, B. Harch, M. Kennard, and F. Sheldon. *In Press*. Integration of science and monitoring of river ecosystem health to guide investments in catchment protection and rehabilitation. Freshwater Biology.

Burnham, K. P., and D. R. Anderson. 1998. Model selection and inference: a practical information-theoretic approach. Springer-Verlag, New York, New York, USA.

Chiles, J., and P. Delfiner. 1999. Geostatistics: Modeling spatial uncertainty. John Wiley and Sons, New York, New York, USA.

Cressie, N. 1993. Statistics for Spatial Data, Revised Edition. Page 900 p. John Wiley and Sons, New York, New York, USA.

EHMP (Ecosystem Health Monitoring Program). 2001. Design and Implementation of Baseline Monitoring (DIBM3): Developing an ecosystem health monitoring program for rivers and streams in southeast Queensland, Report to the Southeast Queensland Regional Water Quality Management Strategy. Brisbane, Australia.

ESRI (Environmental Systems Research Institute). 2006. ArcGIS: Release 9.2 [software]. Redlands, California: Environmental Systems Research Institute.

Harris, J. H. 1995. The use of fish in ecological assessments. Australian Journal of Ecology 20:65–80.

Helsel, D. R., and R. M. Hirsch. 1992. Statistical Methods in Water Resources. Elsevier Science Publishing, New York, New York, USA.

Kennard, M. J., A. H. Arthington, B. J. Pusey, and B. D. Harch. 2005. Are alien fish a reliable indicator of river health? Freshwater Biology 50:174–193.

Kennard, M. J., B. D. Harch, B. J. Pusey, and A. H. Arthington. 2006*a*. Accurately defining the reference condition for summary biotic metrics: a comparison of four approaches. Hydrobiologia 572:151–170.

Kennard, M. J., B. J. Pusey, A. H. Arthington, B. D. Harch, and S. J. Mackay. 2006*b*. Development and application of a predictive model for freshwater fish assemblage composition to evaluate river health in eastern Australia. Hydrobiologia 572:33–57.

Kennard, M. J., B. J. Pusey, B. H. Harch, E. Dore, and A. H. Arthington. 2006*c*. Estimating local stream fish assemblage attributes: sampling effort and efficiency at two spatial scales. Marine and Freshwater Research 57:635–653.

Littell, R. C., R. C. Milliken, W. W. Stroup, and R. Wolfinger. 1996. SAS System for Mixed Models, Cary, North Carolina: SAS publishing.

Martin, R. D. 1980. Robust estimation of autoregressive models. Pages 228–254 *in* D. R. Brillinger and G. C. Tiao, editors. Directions in times series: Proceedings of the IMS special topics meeting on time series analysis. Institute of Mathematical Statistical, Haywood, California, USA.

Moreton Bay Waterways and Catchments Partnership. 2005. South East Queensland Streams and Catchments Version 2. Moreton Bay Waterways and Catchments Partnership, Brisbane, QLD, Australia.

Pusey, B. J., A. H. Arthington, and M. G. Read. 1993. Spatial and temporal variation in fish assemblage structure in the Mary River, South-Eastern Queensland – the influence of habitat structure. Environmental Biology of Fishes 37:355–380.

QNRW (Queensland Natural Resources and Water). 2000. Southeast Queensland 25 meter Digital Elevation Model. Queensland Natural Resources and Water, Indooroopilly, QLD, Australia.

R Development Core Team (2008) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Theobald, D. M., J. Norman, E. Peterson, and S. Ferraz. 2005. Functional Linkage of Watersheds and Streams (FLoWs): Network-based ArcGIS tools to analyze freshwater ecosystems. ESRI User Conference 2005. Environmental Systems Research Institute, Inc., San Diego, California, USA.

Theobald, D. M., J. B. Norman, E. E. Peterson, S. Ferraz, A. Wade, and M. R. Sherburne. 2006. Functional Linkage of Waterbasins and Streams (FLoWS) v1 User's Guide: ArcGIS tools for Network-based analysis of freshwater ecosystems. Natural Resource Ecology Lab, Colorado State University, Fort Collins, CO. Available at http://www.nrel.colostate.edu/projects/starmap/flows_index.htm

Van Rossum, G., and F. L. Drake, Jr. 2005. *The Python Language Reference Manual*. Network Theory, Ltd., Bristol, UK.

Verbeke, G., and G. Molenberghs. 2000. Linear mixed models for longitudinal data. Springer, New York, New York, USA.

Ver Hoef, J. M., and E. E. Peterson. *In press*. A moving average approach for spatial statistical models of stream networks. Journal of the American Statistical Association.

Vogel, R.M., I. Wilson, and C. Daly. 1999. Regional Regression Models of Annual Streamflow for the United States. Journal of Irrigation and Drainage Engineering May/June:148–157.

TABLE B1. Watershed and site-scale explanatory variables considered in the model selection procedure.

Explanatory Variable	Scale	Units	Source
Mean slope	Watershed	%	QNRW 2000
% Conservation	Watershed	%	BRS 2002
% Urban	Watershed	%	BRS 2002
% Mining	Watershed	%	BRS 2002
% Water	Watershed	%	BRS 2002
% Timber production	Watershed	%	BRS 2002
% Agriculture	Watershed	%	BRS 2002
EHMP region	Site	NA	EHMP 2001

TABLE B2. Fixed-effects estimates for the final mixture model (exponential TU/linear-with-sill TD). Parameters were estimated using restricted maximum likelihood (REML).

Effect	**Estimate**	**Std.Error**	**df**	**t-value**	**p-value**
Intercept	0.6669	0.0792	84	8.4209	0
Mean slope	0.0094	0.0061	84	1.5391	0.1275

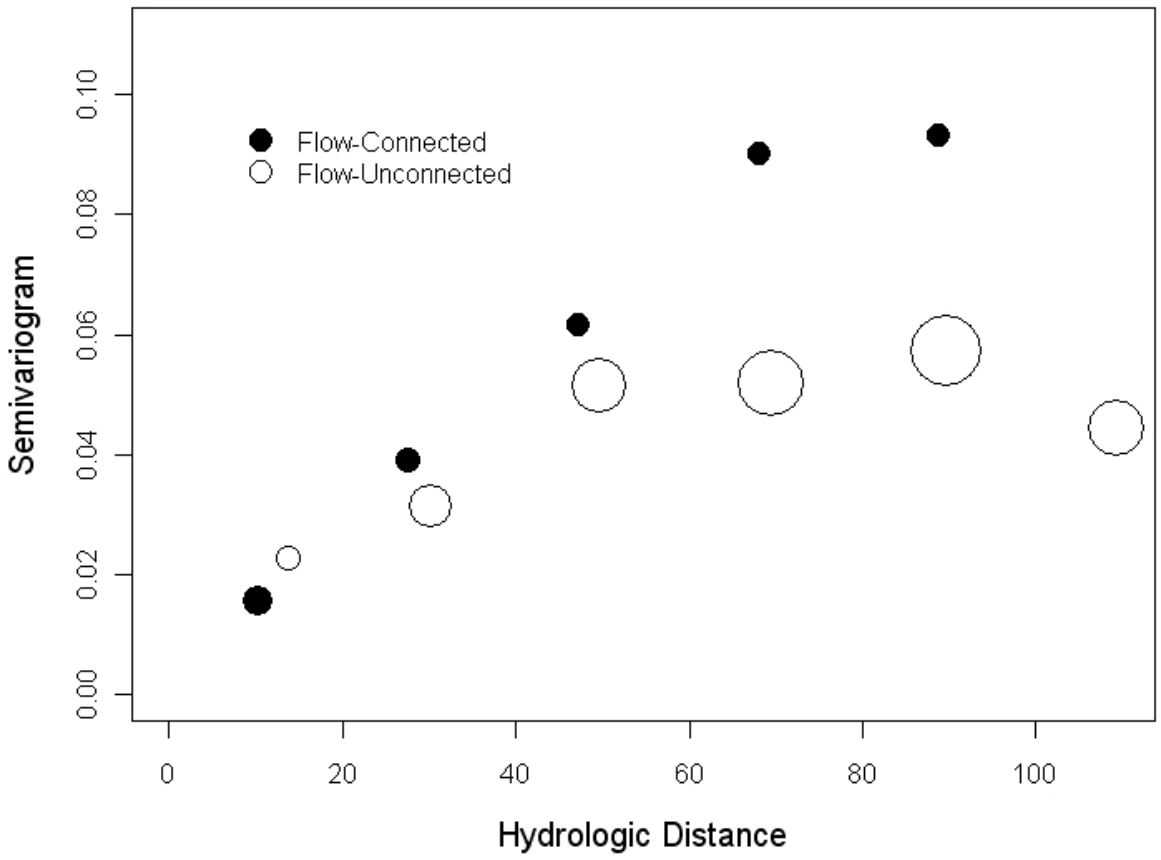

FIG. B1. Semivariogram of the final proportion of native species expected (PONSE) model (exponential TU/linear-with-sill TD) residuals, which includes the hydrologic distance (km) between pairs of sites based on flow-connected (black circles) and flow-unconnected (white circles) relationships. Only lags with > 10 pairs are shown and the size of the circles is proportional to the number of data pairs that are averaged for each value.

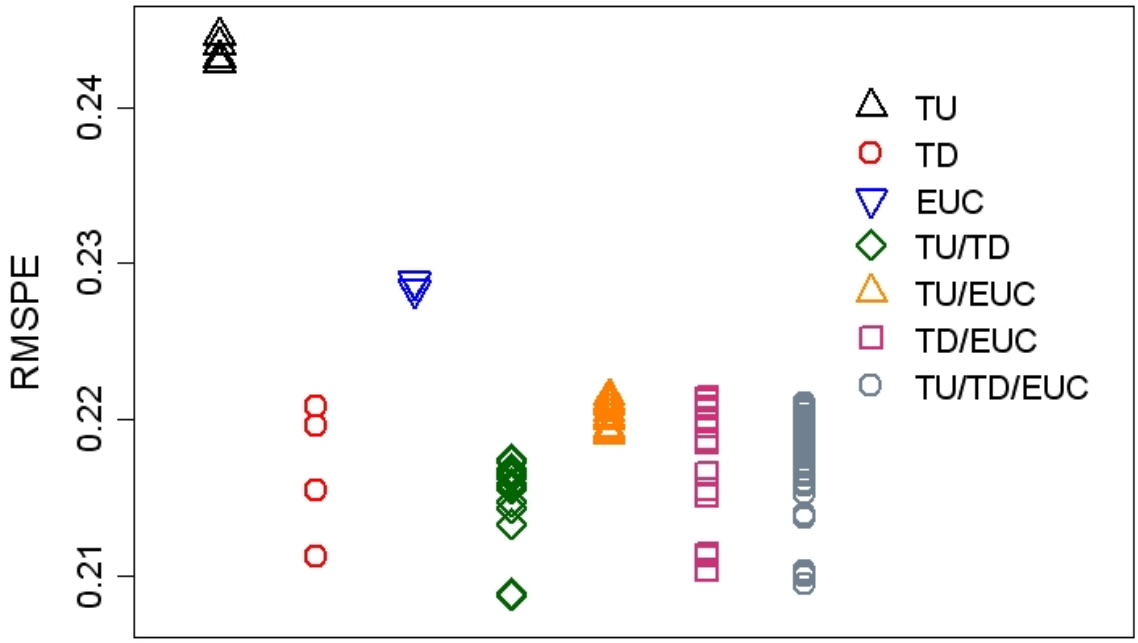

FIG. B2. Root mean square prediction error (RMSPE) values, by mixture type, which were used to select the covariance mixture during the second phase of model selection.