Additional Multivariate Exercise 17

Part 1: Commune level analysis

Data: The data files include 2297 rows, corresponding to 2297 communes rather than 2296, as described in the VER2 textbook; the reason for this discrepancy is not clear. The files with case and control data contain the indicators for single communes, so that each commune is indicated as either a case or a control. There are 446 cases (AI outbreaks). The files do not have a special separator symbol, so the Field Separator must be changed to "Whitespace". The coordinates are Cartesian, not latitude and longitude. The SaTScan results mention that several location IDs were combined into one location; this may indicate a problem with the geographical information for those locations, but we will have to work from the files at hand.

First analysis: The VER2 textbook settings are the defaults: maximal 50% area coverage, only high risk clusters and circular shapes. The results are indeed identical to those of Example 26.8: two clusters with $P < 0.05$:

1) a large cluster with 247 cases among 685 communes, corresponding to a relative risk of 2.92 and a very small P-value; the map shows that a large portion of the circle falls below the study area,
2) a very small cluster with 7 communes that are all cases, corresponding to a relative risk of 5.22 and $P = 0.033$; this cluster is in a different part of the region but also seems to expand a bit beyond the study area,
3) another small cluster with 11 cases out of 15 and $P = 0.063$; this cluster is separated from the two other clusters and fully within the study area.

The main message from this analysis is clearly that a large area in the south of the region has an elevated occurrence of AI outbreaks.

The analysis with elliptic contours is quite a bit slower to run, but it returns exactly the same two circular clusters. The P-values are not exactly the same, most importantly because the class of shapes explored is much wider. Therefore, among the simulated configurations it is *easier* to get clusters with high values for the likelihood-ratio statistic. The P-value for the second cluster is now reported as 0.099. Considering how poorly the circles fit to the study area, it seems reasonable to rely on elliptic rather than circular shapes.

If we identify both high and low risk areas (still with elliptic shapes), the high risk cluster remains unchanged, but a second low risk elliptic cluster appears:

4) a large cluster with 30 cases among 487 communes, corresponding to a relative risk of 0.27 and a very small P-value; the map shows that this cluster is located to the north of the high risk cluster with only minimal overlap and a moderately elliptic shape; also this ellipsis is partly beyond the study area.

Finally, an analysis with Gini Optimized Cluster Collection (in addition to creating clusters hierarchically, as we have been doing so far), provides a more complex output. The Gini clusters may overlap with the hierarchical clusters, and this shows clearly. The SaTScan User Guide is rather vague on how the Gini index is used to select non-overlapping clusters, beyond a broad statement that the Gini index is a measure of statistical dispersion, and "With this criterion, SaTScan selects the group of non-overlapping clusters that maximizes the Gini index, so that there is a big difference in rates between the cluster and non-cluster areas". The previous high and low risk clusters are still included, but we have one additional high risk cluster (essentially a subset of the large circular cluster) and

three additional low risk clusters, two of which overlap considerably with the original low risk elliptic cluster. These new clusters are probably of interest as well and may aid in the interpretation of the results. If we are less concerned about formal significance and consider the analysis as mostly exploratory, there is no reason we cannot include those clusters in our interpretations.

Part 2: District level analysis

Data: After exclusion of districts with a missing value of `duckdens`, a total of 103 districts remain with a total of 364 AI outbreaks in 2006 communes. The Poisson regression analysis in Stata (with the number of communes as an exposure variable) shows the predictor `duckdens` to be highly significant ($\hat{\beta}_1 = 0.00176,\ \mathrm{SE}(\hat{\beta}_1) = 0.00024$), whereas `avg_height` was not quite significant at $P = 0.13$ ($\hat{\beta}_2 = -0.00036,\ \mathrm{SE}(\hat{\beta}_2) = 0.00024$). We therefore try the model with both predictors and compare with the estimation without any predictors. The predicted counts from the two Poisson models are output in suitable variables in the dataset, to be used as expected population counts in SaTScan.

First analysis, without predictors: We use the defaults from the previous analyses, except that we allow for elliptical cluster shapes and include both high and low risk areas from the onset (but no overlapping clusters). The results include one high risk and one low risk cluster:

1) a fairly large high risk cluster with 204 cases among 31 districts (totalling 204 communes), corresponding to a relative risk of 2.88 and a very small P-value; the map shows the cluster as elliptical to the south of the region, and with a substantial part falling outside the study area,
2) a fairly large low risk cluster with 26 cases among 26 districts (totalling 494 communes), corresponding to a relative risk of 0.24 and a very low P-value; this cluster is to the north of the first cluster, essentially non-overlapping and of similar shape.

Despite the first cluster being smaller and of elliptical shape, these two clusters are quite similar to those for the commune level analysis.

Second analysis, with predictors: Still using the same settings, we get these results:

1) a fairly large low risk cluster with 9 cases among 21 districts (totalling 363 communes), corresponding to a relative risk of 0.17 and a low P-value; despite being positioned somewhat differently, this cluster roughly matches the low risk cluster without predictors,
2) a small low risk cluster with 0 cases among 6 districts (totalling 133 communes), corresponding to a relative risk of 0 (obviously!) and a low P-value; this cluster complements the first low risk cluster to cover roughly the area of the large low risk cluster in the previous analysis,
3) a fairly large high risk cluster with 57 cases among 13 districts (totalling 219 communes), corresponding to a relative risk of 2.83 and a low P-value; the map shows the cluster as elliptical to the north-west in the region (where we have not had a high risk cluster before) with a substantial part falling outside the study area,
4) a very small low risk cluster with 1 case among 2 districts (totalling 64 communes), corresponding to a relative risk of 0.06 and $P = 0.0055$,
5) a small high risk risk cluster with 26 cases among 3 districts (totalling 56 communes), corresponding to a relative risk of 3.21 and $P = 0.016$,
6) another small high risk risk cluster with 18 cases among 4 districts (totalling 40 communes), corresponding to a relative risk of 3.55 and $P = 0.034$.

Although the first two low risk clusters taken together are somewhat comparable to what was previously determined, these results are overall quite different. The previous high-risk cluster is no longer there, nor has it been split into smaller high risk clusters. It seems that accounting for the duck density to a large extent has explained the high risk cluster detected previously. We confirm this interpretation by computing the average duck density inside and outside of the high risk cluster without predictors: the values are 388 and 104, respectively.