Free group and project work

Task 4.1: (Discussing your results and preparing for the presentation)

(a) Within your new group of four, present your results from Sheet 3 to the other minigroup (maybe 5 minutes each). Discuss and compare together the results and try to identify why some results in one group are possibly different than in the other. Help each other to understand all steps/results if necessary.

(b) Read Tasks 4.2 to 4.5 below and jointly decide which data set you will continue to work with. Discuss the tasks and possible models/modelling strategies together. Note that 4.2 and 4.3 are relatively "guided" projects, while 4.4 and 4.5 are rather "research-like" (and have not been fully worked through by the tutors). However, please don't hesitate to deviate from the instructions (and do something completely differently!) even if you go with 4.2 and 4.3. **Each data set can only be worked on by at most two groups on a first-come first-serve basis. Please confirm with a tutor that your project is still available!**

(c) The fact that you are working in a group of four does not necessarily imply that one person sits at the computer with the other three around her/him - you can work in parallel, but you just have to make *one* single presentation out of it in the end (which can be presented by one or more members of your group). You now have the possibility to distribute the work among yourselves. You can certainly also agree to examine more than one data set, but to present only (the more interesting) one in the end!

(d) In either case, prepare a short (10 minutes excl. discussion) presentation on your results. The easiest way of doing this is probably Powerpoint. Figures can be integrated e.g. as *.bmp files, which can directly be generated from R graphics.

(e) Note that on Friday CG 66 is booked from 9am on, if you want to start earlier. **Presentations for Group I to VI take place from 1.45pm on in CM 221.**

(f) If you need any literature, please either contact the tutors for a copy or get them from online (durham library electronic access).

Task 4.2: (If you choose the Galaxy Data....)

The main task is to find a suitable model for `velocity` given `north.south` and `east.west`.

(a) Look at additive as well as bivariate models. Fit the models and visualize the results (consider the use of R package **rgl** for the visualization, which is really cool!).

(b) In particular, fit a bivariate local *linear* model generalizing the code provided in the lecture. Visualize the fitted bivariate function. You may need to cut off boundary values.

(c) If not already done, it is interesting to fit a spatial ("geoadditive") model using **SemiPar** and observe where it places the knots.

(d) The relationship between `velocity` and `angle` is also interesting. Provide an adequate graphical summary of their relationship.

(e) You may also want to fit 3D-principal curves through individual slots, or use 2D- principal curves to describe the relationship between pairs of variables.

(f) *Using any suitable model that you have found during your analysis, predict the radial velocity at the origin of Galaxy NGC7531!*

(g) ⋮

Task 4.3: (If you choose the Fetal Data....)

(a) Obtain smooth estimates of NATMOR, NO2, SO2, and CO over time (`DAY`) using an appropriate smoother. Modify the bandwidth or use automatic criteria until the fits look nice. Inspect the shape of the curves and look for similarities.

(b) Now substitute the temporal indicators for `MONTH` in your CORE-model by one single smooth curve `s(DAY)`. Do we still need the `WEEK` indicators then?

(c) Add subsequently (one by one) the pollutants to your model, observe the improvement of goodness-of-fit (in terms of deviance `$dev`) and inspect the fitted curves visually. Are their any strange features?

(d) Decide for a final model, using e.g. information provided by the AIC criterion `$aic`.

(e) Fit your final model also in **BayesX** and compare the results. For the smooth terms included (except `DAY`), consider the extension to *variable* smoothing parameters, substituting `psplinerw2` by `tpsplinerw2`.

(f) *Using any suitable model that you have found during your analysis, predict the number of intrauterine mortalities in the city of São Paulo on 27th of January 1991 (this corresponds to one of the rows initially omitted). Use e.g.* `fetal[27,1:12]` *to extract the necessary predictor data. If your chosen model involves the missing value CO, interpolate it from the data on 26th and 28th of January. Compare your result with the observed response.*

(g) ⋮

Task 4.4: (If you choose the Age/Income Data....)

(a) Implement the bootstrap methodology outlined in the notes in order to construct bootstrap confidence intervals for the age/income data.

(b) Compare your results with pointwise and simultaneous confidence intervals obtained via `locfit` and BayesX.

(c) Try to improve on your bootstrap confidence intervals, using methods such as outlined in

- W. Hardle and A.W. Bowman (1988), Bootstrapping in nonparametric regression: Local adaptive smoothing and confidence bands. *Journal of the American Statistical Association* 83, 123–127. and/ or
- T.L. McMurry and D.N. Politis (2008), Bootstrap confidence intervals in nonparametric regression with built-in bias correction. *Statistics and Probabbility Letters* 78, 2463–2469.

(d) ⋮

Task 4.5: (If you choose the Zambia Data....)

(a) Carry out an adequate analysis of the Zambia data **in R** and compare with your results from the BayesX session.

(b) We cannot use the map-based information in R (at least, not directly), but **SemiPar** should allow you to set up a spatial random effect (option `random=...`). This does not account for between-district-correlation, but *within-district-correlation*, giving the same intercept to all individuals within one region. You should read the SemiPar manual in order to find out how to do this.

(c) Actually, such a random effect model can also be fitted in **BayesX**, using the additive component `district(random)` ("unstructured spatial effect" in BayesX terminology). Try this and compare your results to (b). If you need help with this subquestion consult the second reference provided on Handout 1.

(d) ⋮