UNIVERSITY OF CAMBRIDGE

MATHEMATICAL TRIPOS Part III

Thursday, 1 June, 2017 9:00 am to 12:00 pm

PAPER 206

APPLIED STATISTICS

*Attempt no more than **FOUR** questions.*

*There are **SIX** questions in total.*

The questions carry equal weight.

STATIONERY REQUIREMENTS
Cover sheet
Treasury Tag
Script paper

SPECIAL REQUIREMENTS
None

You may not start to read the questions printed on the subsequent pages until instructed to do so by the Invigilator.

1

(a) Let Z be a random variable with distribution belonging to an exponential dispersion family with natural parameter θ and dispersion ϕ. Derive the expression of the expected values $\mathbb{E}[Z]$ and the variance $\mathrm{Var}(Z)$ as functions of θ and ϕ.

(b) Let Y be a negative binomial random variable with density function

$$f_Y(y; r, \lambda) = \frac{\Gamma(y+r)}{y!\Gamma(r)}\Big(\frac{\lambda}{r+\lambda}\Big)^y\Big(\frac{r}{r+\lambda}\Big)^r, \quad y \in \{0, 1, 2, \ldots\},$$

with $r > 0$ a known parameter. Show that Y belongs to an exponential dispersion family and identify the natural parameter and the dispersion parameter. Using the expressions derived in part (a), compute the mean and the variance of Y.

The administrators of a fishing reservoir are investigating if the daily number of fishes caught by their customers is changing over time. They imported in R the data from the most recent year about the number of fishes caught by each client (`nfishes`) and the day they visited the reservoir (`day`, ranging from 1 to 365). They fitted the following model:

```
> library(MASS)
> fish.model<-glm.nb(nfishes~day)
> summary(fish.model)
```

```
##
## Call:
## glm.nb(formula = nfishes ~ day, init.theta = 195525.8785, link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -5.6444  -1.4817   0.3818   1.3630   4.4604
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  2.778e+00  1.605e-02  173.12   <2e-16 ***
## day         -3.169e-03  8.857e-05  -35.79   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(195525.9) family taken to be 1)
##
##     Null deviance: 8868.0  on 1299  degrees of freedom
## Residual deviance: 7544.3  on 1298  degrees of freedom
## AIC: 11808
##
## Number of Fisher Scoring iterations: 1
```

(c) Write down the algebraic form of the fitted model and the estimates of the parameters. If the assumptions of the model are valid, what can you conclude about the change in the number of fishes caught over time?

(d) The administrators are worried that some of the visitors came to the reservoir only to keep company to their friends who were fishing, while not really trying to catch any fish themselves. Indeed the number of zeros in the data was much larger than would be expected from the fitted model. Propose a new model to address this problem.

(e) Write down the likelihood of this new model.

(f) Describe how this model can be fitted with an EM algorithm, specifying the expressions of the expectation and maximization steps for this specific model (you can assume r known).

2

A company is investigating the probability that users of their online store make a purchase. This probability may depend on the price of the product and on the layout of the website. They collected data about individual users making a purchase or not (`purchase`, coded as 1 if the user made the purchase, 0 else), the price of the product they were considering (`price`) and the layout of the website (`layout`, a factor with levels "A", "B" and "C" to indicate the three possible layouts). Consider the following (edited) R output.

```
> model1<-glm(purchase~price+layout,family="binomial")
> model2<-glm(purchase~price,family="binomial")
> anova(model1,test="Chisq")

## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: purchase
##
##        Df Deviance  Resid. Df  Resid. Dev   Pr(>Chi)
## NULL                      119     148.263
## price   ?    ?            118     101.920   9.93e-12
## layout  ?    ?             ?       97.571     0.1136

> summary(model2)

##
## Call:
## glm(formula = purchase ~ price, family = "binomial")
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6689  -0.6456  -0.3200   0.6891   2.4352
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  5.16258    1.13031   4.567 4.94e-06 ***
## price       -0.08207    0.01595  -5.147 2.65e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
```

(a) Write down the algebraic form of the models fitted in `model1` and `model2`. Derive the expression of the deviance for the two models.

(b) Determine the value of the five numbers that have been substituted by question marks in the Analysis of Deviance Table. Which model is preferable? Describe the hypothesis test that has been carried out to reach this conclusion.

(c) Write down the estimates and 95% confidence intervals for the parameters of the chosen model (recall that the 0.975 quantile of a standard normal distribution is $Z_{0.025} = 1.96$). What can you conclude about the relationship between layout, price and the probability of purchase? What is the predicted probability of purchase for a product of price 100 with layout "B"?

(d) A third model is fitted with the commands

```
> model3<-glm(purchase~price,family = "quasibinomial")
> summary(model3)
##
## Call:
## glm(formula = purchase ~ price, family = "quasibinomial")
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6689  -0.6456  -0.3200   0.6891   2.4352
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.16258    1.08921   4.740 6.03e-06 ***
## price       -0.08207    0.01537  -5.341 4.55e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Describe the model being fitted as `model3` and how its parameters are estimated. Is this an improvement with respect to `model2`?

3

(a) Define the space of *natural cubic splines* on the set of knots $X_1 < \cdots < X_K$ and compute its degrees of freedom.

Consider the model

$$Y_i = m(X_i) + \epsilon_i, \quad \text{for } i = 1, \ldots, n,$$

where $\epsilon_1, \ldots, \epsilon_n$ are independent $N(0, \sigma^2)$ random variables.

(b) Describe how the unknown regression function m can be estimated by regression splines, using a cubic natural splines basis on the knots $X'_1 < \cdots < X'_K$, and find the expression of the estimator. Discuss the role played by the number and the positions of the knots and possible strategies to select both.

(c) Describe how the unknown regression function m can be estimated by cubic smoothing splines for a fixed value of the smoothing parameter λ and find the expression of the estimator.

Researchers are investigating the relationship between the number of prescriptions in a year (`npres`), age (`age`) and the white blood cell counts (`wbc`) in patients. They fit the following model:

```
> library(mgcv)
> model1<-gam(npres~s(age,bs="cr")+s(wbc,bs="cr"))
> plot(model1)
```

(d) Write down the algebraic form of the fitted model. Looking at the estimated regression functions in Fig.1, comment on how the number of prescription is related to age and white blood cell count. Suggest how this model can be improved and write down the R commands to fit the new model.

The researchers later realized that this dataset has been generated by putting together the information coming from four different doctors. They then decide to use a factor (`doctor`, taking values 1,2,3 and 4) to indicate which doctor treated each patient in the dataset and to fit the new model

```
> model2<-gamm(npres~s(age,bs="cr")+s(wbc,bs="cr")+ random=list(doctor=~age))
```

(e) Write down the algebraic form of `model2` and explain why the researchers fitted this new model.

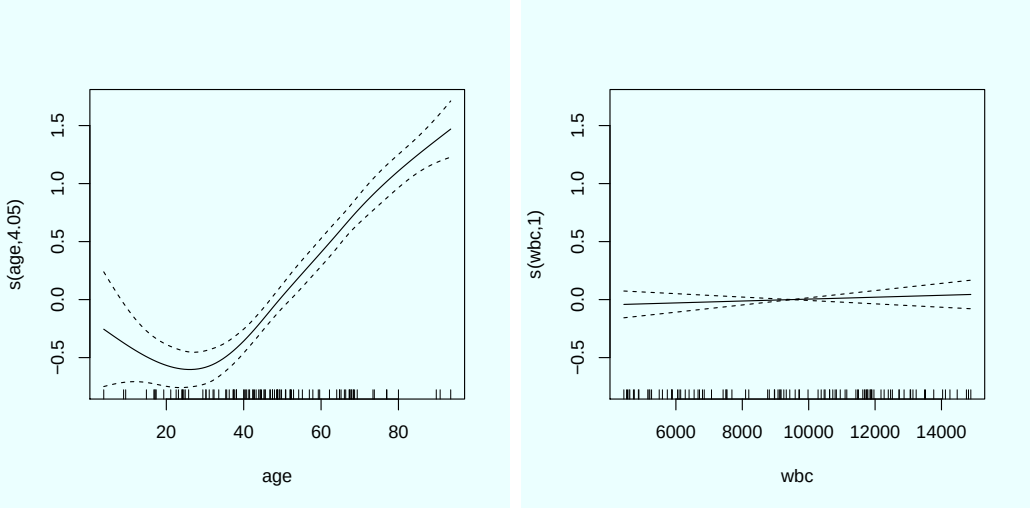

Figure 1: Estimated regression functions for `model1`.

4

Data have been collected to study the speed of growth of blowfly larvae. The size of the larvae (`size`, in mm) and the hours since hatching (`hours`) have been imported in R. After hatching, larvae have been randomly assigned to one of four different incubators and this is denoted in R with a factor `incubator` with levels `A`,`B`, `C` and `D`. The following analysis has been carried out in R.

```
> library(lme4)
> fly.model<-lmer(size~hours+(0+hours|incubator))
> summary(fly.model)

## Linear mixed model fit by REML ['lmerMod']
## Formula: size ~ hours + (0 + hours | incubator)
##
## REML criterion at convergence: 14041.1
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.3551 -0.6053  0.1053  0.7817  1.9734
##
## Random effects:
##  Groups    Name  Variance  Std.Dev.
##  incubator hours   0.00642  0.08013
##  Residual        374.64823 19.35583
## Number of obs: 1600, groups:  incubator, 4
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept) 24.32646    0.97143   25.04
## hours        0.70005    0.04093   17.10
##
## Correlation of Fixed Effects:
##       (Intr)
## hours -0.178
```

(a) Write down the algebraic form of the fitted model and the estimates of the parameters. Explain why it may be sensible to model the effect of the incubator as random and why a random intercept is not needed in the model. What is the best prediction for the size of a larva after 10 hours from hatching in a new incubator (not present in the original study)?

(b) Derive the expression of the likelihood ratio test statistics to test if the random effect is needed in the model. Explain why it is not possible to approximate its distribution with a χ^2 distribution in this case and describe an alternative procedure to carry out the test.

(c) State the definition of the AIC and derive its expression for `fly.model`.

(d) Let us assume now that, *before hatching*, blowfly eggs have been collected from three different locations and these locations have been recorded in the `location` factor. Modify the model to take into account the variability in size due to the different locations and write down the R code to fit this new model.

(e) An entomologist questioned the linear relationship between the size and the hours after hatching. Consider the following R commands.

```
> library(mgcv)
> fly.model2<-gamm(size~s(hours,bs="cr"), random=list(incubator=~0+hours))
```

Describe the model fitted in `fly.model2` and explain which figure can be plotted in R to investigate if it is reasonable to assume a linear relationship between the size and the hours after hatching.

5

(a) Let $Y_1, \ldots, Y_n$ be a sample from a weakly stationary process with mean μ and autocovariance function $\gamma(h)$, $h = 0, 1, \ldots$. Show that the sample mean is an unbiased estimator for μ and compute the expression for its variance.

(b) In the same setting of part (a), assume now that $\mu = 0$ and is known. Write down the expression of the sample autocovariance function in this case and compute its expected value.

(c) Define a zero-mean autoregressive process of order 1 and derive the expression for its autocovariance and autocorrelation functions when the process is causal.

(d) The R vector `temperature` contains a time series of daily average temperatures for two months. Consider the following R code and output:

```
> library(forecast)
> plot(temperature,xlab="day")
> acf(temperature)
> model_temp<-auto.arima(temperature,seasonal = FALSE, stationary = TRUE)
> model_temp

## Series: temperature
## ARIMA(0,0,1) with non-zero mean
##
## Coefficients:
##          ma1     mean
##       0.6727  15.0133
## s.e.  0.1148   0.4505
##
## sigma^2 estimated as 4.634:  log likelihood=-132.61
## AIC=271.22   AICc=271.64   BIC=277.55
```

Write down the algebraic form of the model that has been selected by the `auto.arima` function and the estimates for the parameters. Is the choice of the options `seasonal = FALSE` and `stationary = TRUE` justified? You can find the plot of the time series and its the sample autocorrelation function in Fig.1.

(e) Based on the selected model, what is the expression of the predictor for the temperature one week in the future? What is its expected prediction error? Compare this with the expected prediction error for the naive predictor: which one is better?

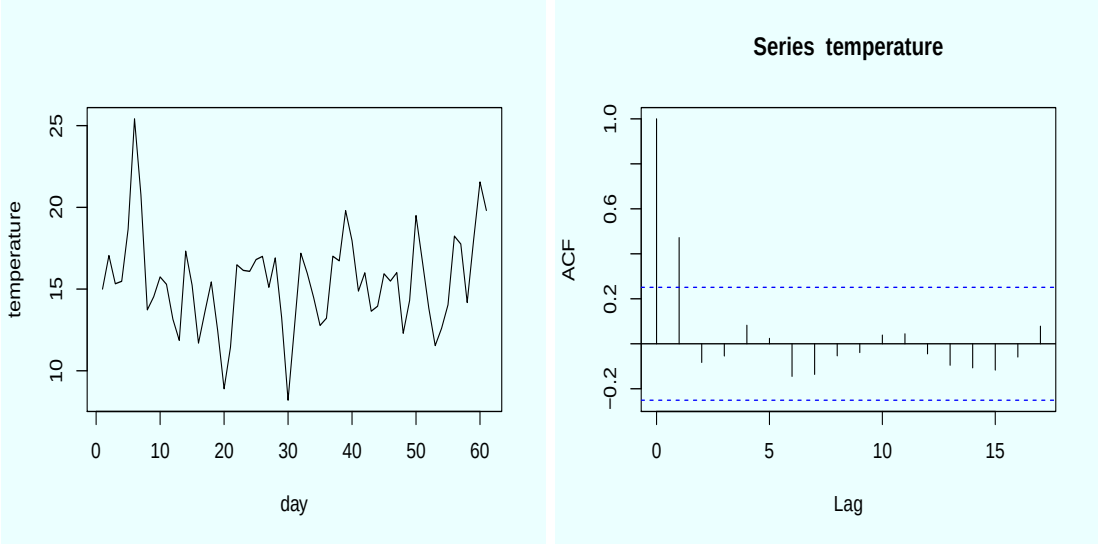

Figure 1: Time series of daily average temperature (left) and its sample autocorrelation function (right).

6

(a) Consider the Gaussian semivariogram model $\gamma : \mathbb{R} \to \mathbb{R}$,

$$\gamma(t) = \begin{cases} \tau^2 + \sigma^2(1 - \exp(-\phi^2 t^2)) & \text{for } t > 0 \\ 0 & \text{else} \end{cases}$$

Show that $\gamma(t)$ defines a second-order stationary process and derive the expression of the corresponding covariogram $C(t)$. Find out the nugget, the sill and the range for this model.

(b) Let $Z_{\mathbf{s}_1}, \ldots, Z_{\mathbf{s}_n}$ be generated from a Gaussian process with zero mean and covariogram $C(t)$. Derive the expression of the best linear unbiased predictor for the value of the field $Z_{\mathbf{s}_0}$ in a new location $\mathbf{s}_0$ and of a 95% prediction interval for $Z_{\mathbf{s}_0}$ (you can denote with q_α the $1-\alpha$ quantile of a standard normal distribution and you do not need to write down the expression of $C(t)$ here).

A maritime agency collected depth data from 52 points in a navigable channel and they are interested in producing an accurate map of the depth in the whole channel for navigation purpose. Consider the following R code and output, together with the graphics in Fig. 1 (the semivariogram model `Gau` corresponds to the Gaussian semivariogram defined in part (a))

```
> spplot(depth_data,main='depth',xlab="x",ylab="y")
> smvg <- variogram(depth ~ 1, width=0.1, data=depth_data)
> gauss.model<-fit.variogram(smvg, vgm(10, "Gau", 2, 1))
> plot(smvg,gauss.model,main="smvg",ylim=c(0,10))
> smvg2 <- variogram(depth ~ y, width=0.1, data=depth_data)
> plot(smvg2,gauss.model2,main="smvg2",ylim=c(0,0.8))
> gauss.model2
##   model     psill    range
## 1   Nug 0.0000000 0.000000
## 2   Gau 0.4726257 1.112235

> gauss.gstat <- gstat(formula = depth ~ y, data = depth_data, model=gauss.model2)

> s0<-data.frame(x=0,y=0,y=0)
> coordinates(s0)<-c("x","y")
> gauss.trend<- predict(gauss.gstat,newdata = s0 ,BLUE=TRUE)
> gauss.trend
##   coordinates var1.pred   var1.var
## 1      (0, 0)  20.13793 0.09456496
> s1<-data.frame(x=0,y=1,y=1)
> coordinates(s1)<-c("x","y")
> gauss.trend1<- predict(gauss.gstat,newdata = s1 ,BLUE=TRUE)
> gauss.trend1
##   coordinates var1.pred   var1.var
## 1      (0, 1)  22.08399 0.06097459
```

(c) Comment on the binned semivariogram in `svgm` and explain why it is needed to fit the new model in `svgm2`.

(d) Write down the algebraic form of the model fitted in `gauss.gstat` and the estimates of the parameters.

(e) Describe the procedure used to estimate jointly the parameters of the drift term and the spatial dependence.

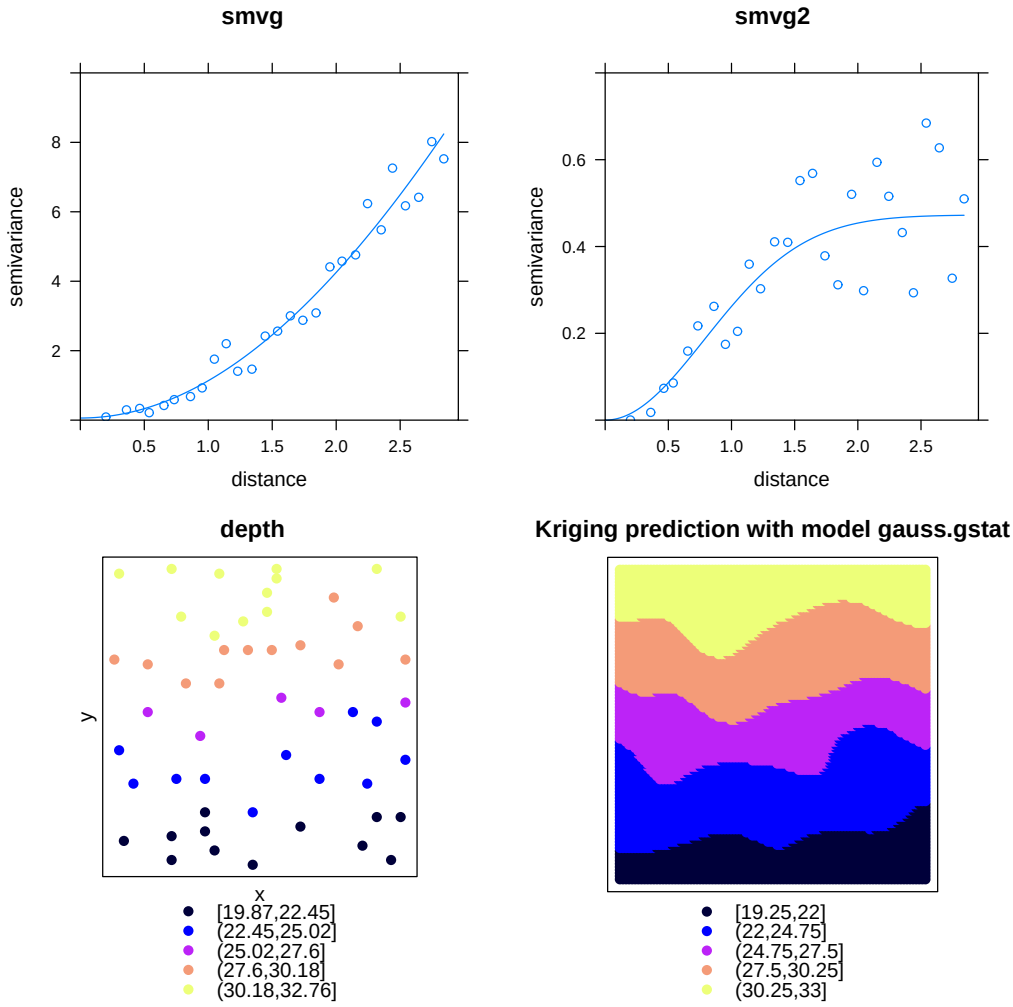

Figure 1: Top left: Binned semivariogram and fitted semivariogram model for `smvg`. Top right: Binned semivariogram and fitted semivariogram model for `smvg2`. Bottom left: Observed depth data. Bottom right: Kriging prediction for the depth field.

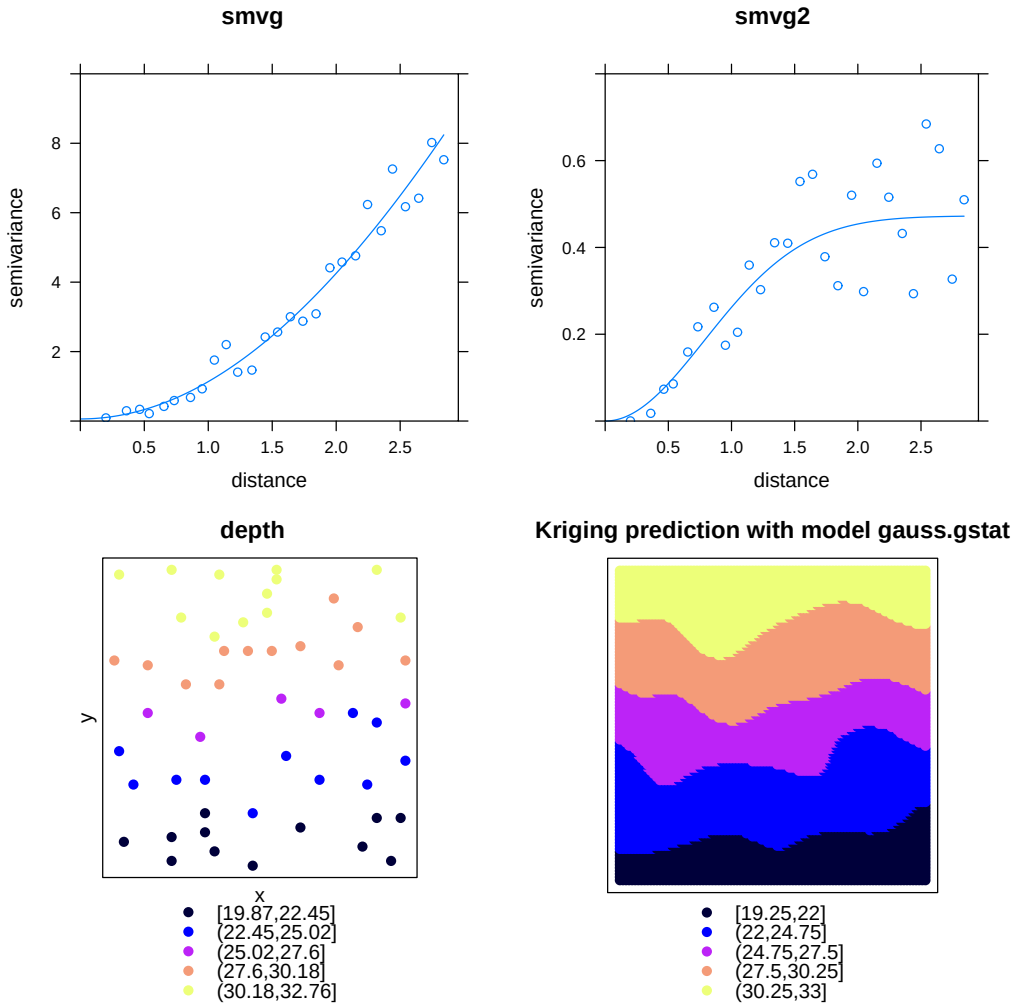

Figure 1: Top left: Binned semivariogram and fitted semivariogram model for `smvg`. Top right: Binned semivariogram and fitted semivariogram model for `smvg2`. Bottom left: Observed depth data. Bottom right: Kriging prediction for the depth field.

END OF PAPER