Assignment

Robusta Metoder
MS1415

Bruna G. Palm
Department of Mathematics and Natural Sciences
Blekinge Institute of Technology, Sweden

April 5, 2024

Introduction

For this assignment, you are to work in groups of two (2) students or alone. The objective of this assignment is to implement different statistical models on provided data sets, aiming at identifying the most suitable one. The result from this assignment should be a short written report (PDF format) that you submit via the course page on Canvas before the deadline (which also is stated on Canvas). You must motivate your steps aiming at validating your conclusions. Additionally, try to write short and to the point.

Grades

This assignment is graded with the following grades: `G/Ux/U`.

Allowed programming language, packages, and tools

Use the language you feel more comfortable with!

Upon submission of your report

Before you upload your report make sure:

1. That you have included your **names** and **email addresses** in the report.
2. The report follows a logical and a well-structured **format** with written explanations.
3. That you have carefully checked the report for **spelling and grammatical errors**.
4. That your report is written in **English**.

Failure to comply with any of the aspects above could result in a failing grade, and that you have to revise the report and submit it again on a later deadline.

Good luck!

Ground Type Detection in a SAR Image
Part 1 — Regression Models

The data set

This part of the assignment, you work with one data set related to a SAR image. Data is available on Carnvas. The data set is sourced from publicly available information and it is fully discussed and presented in Gomez et al. (2017). The employed data set is the San Francisco Bay SAR image and the image is available in https://ctim.ulpgc.es/polsar/. Figure 1 shows the intensity data of the 200 × 350 San Francisco Bay image. The ground scene of the evaluated image is dominated by ocean (dark ground), forest (gray ground), and urban area (light ground).

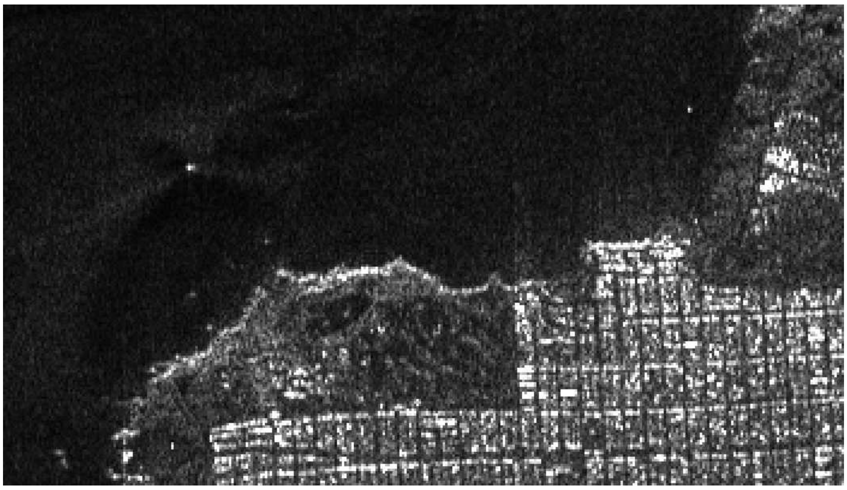

Figure 1: Original San Francisco SAR images.

Detection Theory

Let's assume that the mean of an observed signal (y) presents different values depending on the ground type. To illustrate this idea, consider a region of forest in an image. To detect if this particular area has the same behavior as another one, we need to fit a regression model as follows:

$$g(\mu_i) = \beta_1 + \beta_2 x_{2i} + \sum_{j=3}^{r} \beta_j x_{ji},$$

where (i) β_1 is the intercept; (ii) x_{2i} is a binary covariate equal to one if the region consists of forest and zero otherwise; and (iii) x_{ji}, $i = 3, 4, \ldots, r$, are any other covariates that can influence the mean of y. The detection problem is to distinguish between the hypotheses:

$$\begin{cases} \mathcal{H}_0 : \mu_i = g^{-1}(\beta_1 + \sum_{j=3}^{r} \beta_i x_{ji}), \quad (\beta_2 = 0), \\ \mathcal{H}_1 : \mu_i = g^{-1}(\beta_1 + \beta_2 x_{2i} + \sum_{j=3}^{r} \beta_i x_{ji}). \end{cases}$$

The detector idea is to verify if the null hypothesis is rejected, i.e, $\beta_2 \neq 0$ and the forest land use is detected. This technique can be considered to detect any type of ground in SAR images.

Your Task

Your task is to apply the detection theory presented above in the three different regions in the San Francisco image, i.e., forest, sea, and urban area, considering different regression models (considering different distributions). This idea is illustrated in Figure 2; in particular, the boxes A, B, and C represent the sea, forest, and urban ground types, respectively. To perform the detection, you need to model the mean of the response signal using an intercept ($x_{1i} = 1 \forall i$) and two dummy variables (x_{2i} and x_{3i}) representing each tested region, as

$$g(\mu_i) = \beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i}.$$

The response signal is composed of the vectorized pixel values of the three tested areas (sea, forest, and urban areas). Variable x_{2i} is defined as one for forest ground type and zero for the rest. The variable x_{3i} is defined as one for urban area and zero for the others. The sea region is represented when $x_{2i} = 0$ and $x_{3i} = 0$.

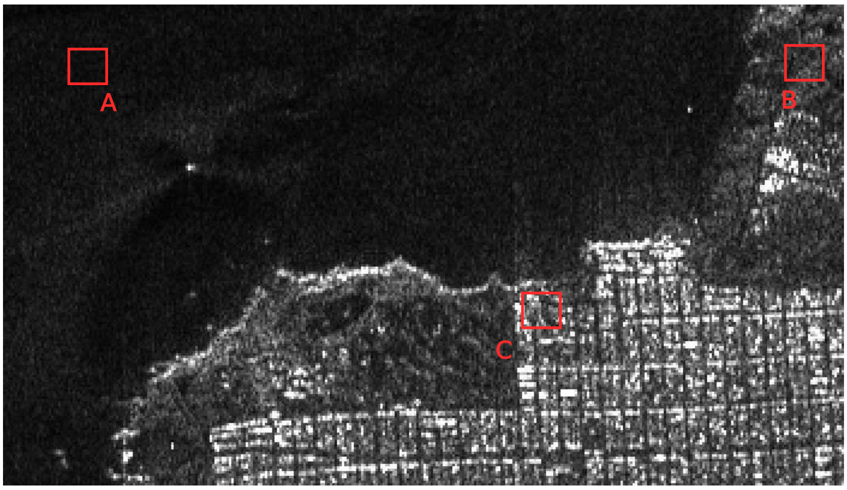

Figure 2: San Francisco SAR image, where the red boxes are related to sea, forest, and urban areas.

Consequently, your task is to fit a GLM based on the Gamma distribution, a model based on the Rayleigh distribution, and a model based on the normal distribution, for example, and try to identify the most suitable one for ground type detection in these particular images. Additionally, verify the quality of the adjusted models and identify the mean response relationship with x_{2i} and x_{3i}. Evaluate the detectors considering two or three different window sizes. Your task is summarized in the following.

1. Define the tested regions;
2. Create the observed signal with the vectorized pixels of these three regions using windows of 20 × 20 pixels;
3. Check the data behavior to verify if the considered regression models are suitable approaches to fit such data;
4. Create two dummy covariates;
5. Fit the selected regression models. You can use the functions available on Canvas.
6. Perform the detection theory. Are the covariates significant to the model? Are they introducing information about variations in y?

7. Test the residuals. Is the model correctly specified? (Consider a residual vs index plot and check evidence of normality with a histogram, for example).

8. Check the relationship between the mean of y and the dummy covariates.

9. Verify the determination coefficient for the fitted models.

10. In conclusion: what is the most accurate model for such data, considering detection and modeling evaluation?

Time Series Modeling and Predicting
Part 2 — Time Series Models

The data set

The data sets available for this part of the assignment refer to the rate of hidden unemployment due to substandard work conditions in a Brazilian city. The data is sourced from publicly available information and it can be found in[1] and is available on the course page on Canvas. Figure 3 show the hidden unemployment rate data set.

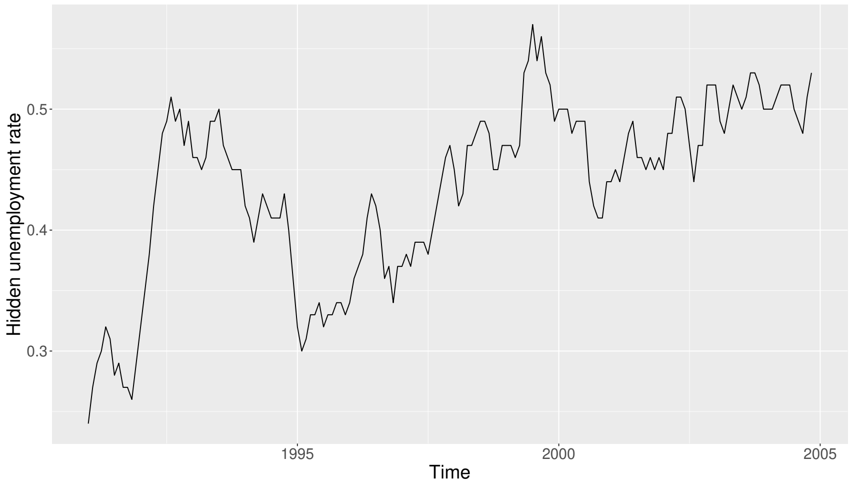

Figure 3: Hidden unemployment rate time series.

Your Task

Your task is to apply time series models to predict the unemployment rate. Considering the characteristic of the data, suitable distributions to fit such data are the beta distribution and the Kumaraswamy distribution, for example. Thus, your task is summarized in the following.

1. Split the data set in to test and validation part, separating the last 10 observed values to evaluate the forecasting;

2. Check the data behavior to identify suitable approaches to fit such data;

[1] See http://www.ipeadata.gov.br

3. Fit the Gaussian-base and model based on suitable distributions. To fit, for example, the beta and Kumaraswamy-based ARMA models (βARMA model and KARMA model, respectively), you can use the functions available on Canvas.
4. Evaluate the fitted models.
5. Test the residuals. Are the models suitable for the tested data sets?
6. Evaluate the prediction capability of the tested models in the test and validation data sets.
7. In conclusion: what is the most accurate model for such data?

References

Gomez, L., Alvarez, L., Mazorra, L., Frery, A. C., 2017. Fully PolSAR image classification using machine learning techniques and reaction-diffusion systems. Neurocomputing 255, 52–60.