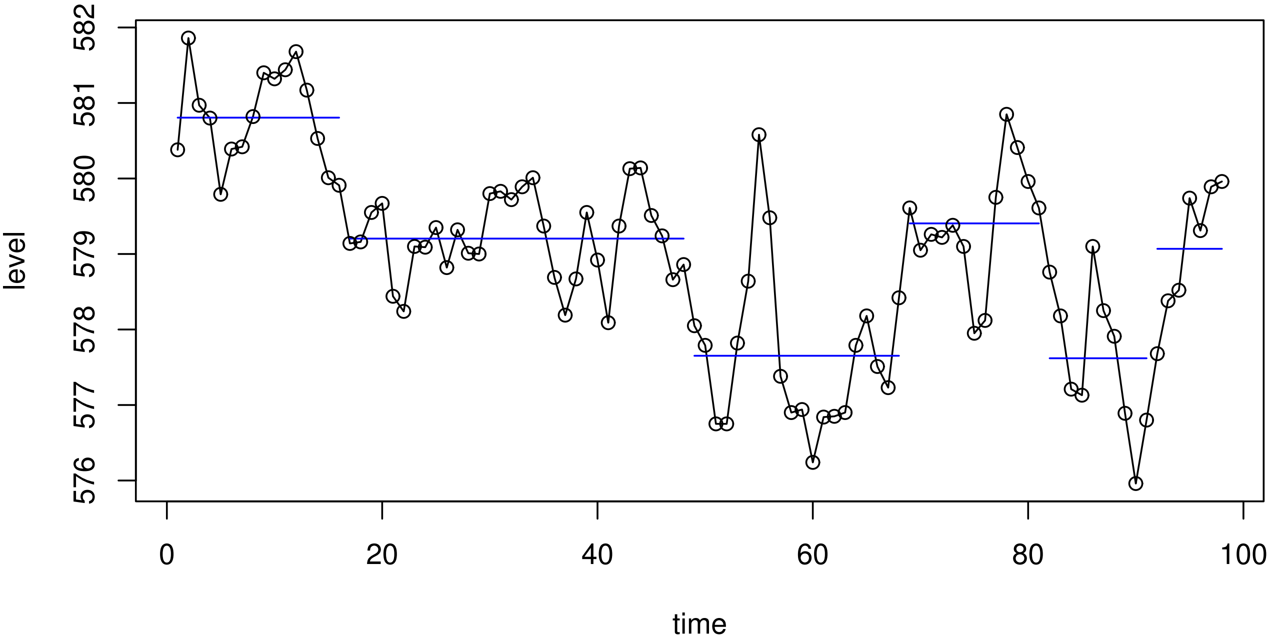
36-350: Data Mining

Homework 14 and Lab 14
Date: November 25, 2002 **Due: end of lab December 6, 2002**

1. In class it was shown how to use Ward's method to find change-points in a time series. A simpler idea for finding change-points is to look for a large difference in the level between consecutive years. Does this idea work on the Lake Huron data? The data is plotted below, along with the result of Ward's method. The individual measurements are shown as circles.



2. In this problem and the next, you will use regression techniques to model a time series. The predictor is time. The file `uspop.rda` gives, in the variable `x`, the population of the United States (in millions) as recorded by the census every ten years from 1790–1970.

   (a) Plot the population versus time. What anomalous years do you see?

   (b) Break the series into pre-1860 and post-1860. If the data frame is x, this can be accomplished by

   ```
   x1 <- x[(x[,"time"] <= 1860),]
   x2 <- x[(x[,"time"] > 1860),]
   ```

   Make a transformation and use linear regression to fit a model for the pre-1860 population as a function of time. Give a plot or two to argue that a linear model fits well under the transformation. State the model as a formula for uspop (the original variable).

   (c) Plot the residuals for your pre-1860 model. Which two years pre-1860 are most unlike the others (have the largest residuals)?

3. (a) Make a transformation and use linear regression to fit a model for the post-1860 population as a function of time. Give a plot or two to argue that a linear model fits well under the transformation. State the model as a formula for uspop (the original variable).

   (b) Plot the residuals for your post-1860 model. Two years are outliers. Which are they?

   (c) Remove the outlier years. You can remove a year as follows:

   ```
   x2 <- x2[(x2[,"time"] != year),]
   ```

   Refit the model and plot residuals. One year is unusually high. Which one? (The page `http://www.missouri.edu/~socbrent/immigr.htm` may be of interest.)

   (d) What does your refitted model predict for the population in year 2000?