

STATS 531
Winter, 2020
Midterm Exam

We investigate a time series on over-crowding in the Emergency Room of the University of Michigan Hospital. The data, $y_{1:N}$ with $N = 24 * 365$, are hourly occupancy fractions for one year, starting July 1st 2005. Occupancy fraction is defined to be the mean number of patients in the ER during each hour divided by the total number of beds available (the ER operates 24 hours a day, 7 days a week, 365 days a year). Note that the occupancy fraction, shown in Fig. 1, can exceed one. The purposes of investigating these data are to predict future occupancy, and to make progress toward relating ER overcrowding with other variables such as errors in medical procedures.

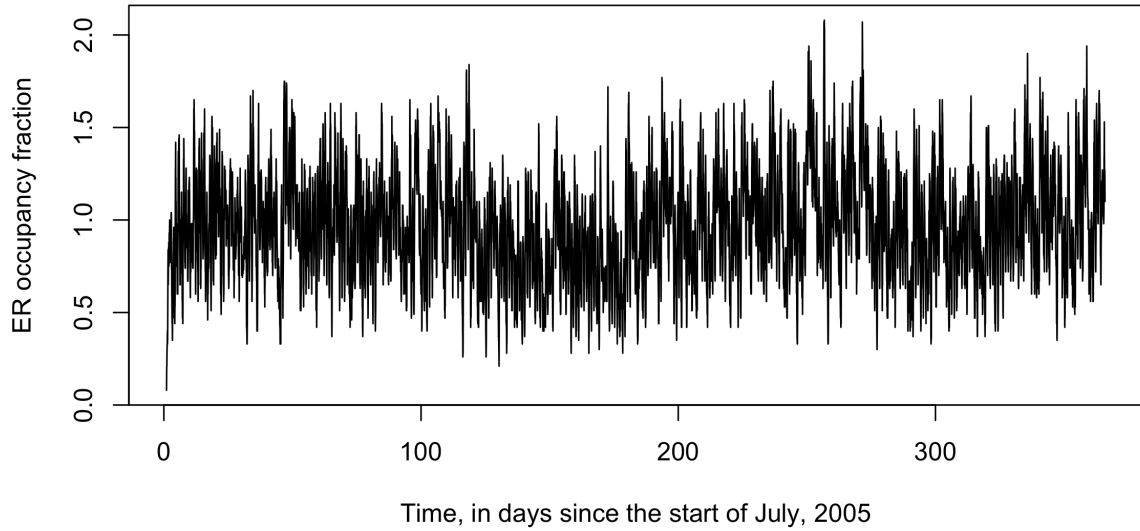


Figure 1: Hourly occupancy fraction at the University of Michigan Emergency Room

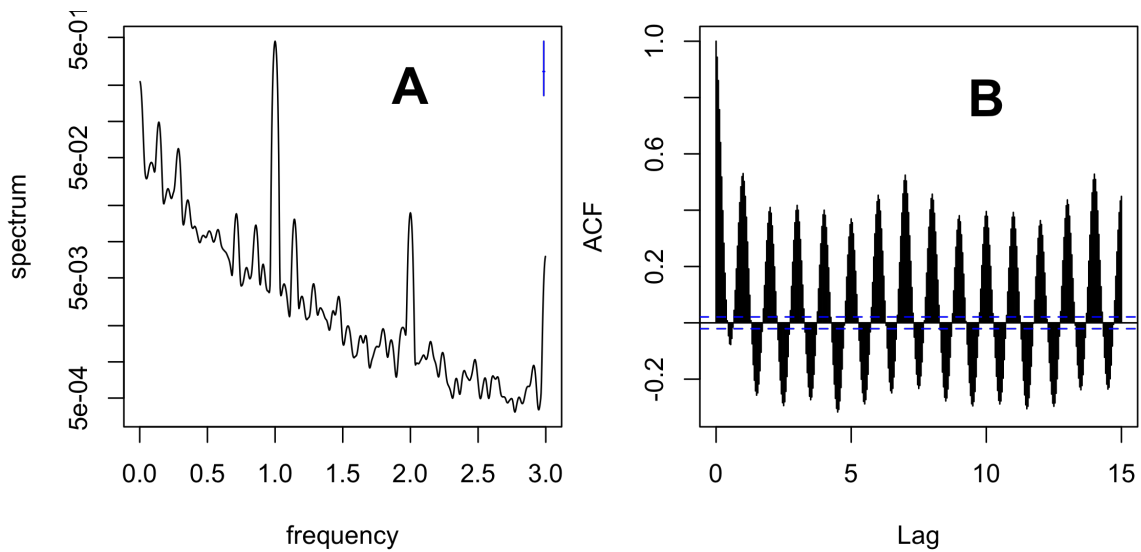


Figure 2: (A) Smoothed periodogram of $y_{1:N}$. (B) sample auto-correlation function of $y_{1:N}$

SECTION A. Fig. 2 shows a smoothed periodogram and an ACF of the data.

A1. [1 point] What are the units of frequency in Fig. 2A? Explain your reasoning. Hint: Care is needed to make allowance for the x-axis truncation pointed out below.

From figure 2B, we see a strong daily cycle and small additional correlation at 7 days (1 week). The strong 1 cycle/day peak is evident at frequency 1 in Fig. 2A, so units must be cycles/day.

A2. [2 points] Explain how you can tell that the periodogram in Fig. 2A has been truncated to exclude high frequencies (this is done to show more clearly the information at lower frequencies).

The highest frequency on a periodogram is 0.5 cycles per observation. This corresponds to 12 cycles per day. The axis has therefore been cut at frequency = 3 cycles per day.

A3. [3 points] Using Fig. 2, can you reject a null hypothesis that there is no weekly pattern to occupancy fraction? Explain. Hint: the bar top right in Fig. 2A may be useful, though the horizontal cross for this bar is small and hard to see.

A confidence interval around the peak at frequency 1/7 cycles/day, constructed using the error bar in Fig. 2A, excludes the base of this peak. Thus, we can reject the null hypothesis that this peak is chance variation.

SECTION B. Fig. 1 suggests that the occupancy could be modeled by a random process $Y_{1:N}$ whose expected value $\mu_n = \mathbb{E}[Y_n]$ is slowly varying with time. The variation around the mean in Fig. 1 appears quite stable. Thus, it may be reasonable to model $y_n - \hat{\mu}_n$ as a stationary process, with $\hat{\mu}_n$ constructed using local regression. This is done here using the R command `mu.hat=loess(y~time,span=0.5)$fitted`. The estimate $\hat{\mu}_t$ of μ_n is shown in Fig. 3.

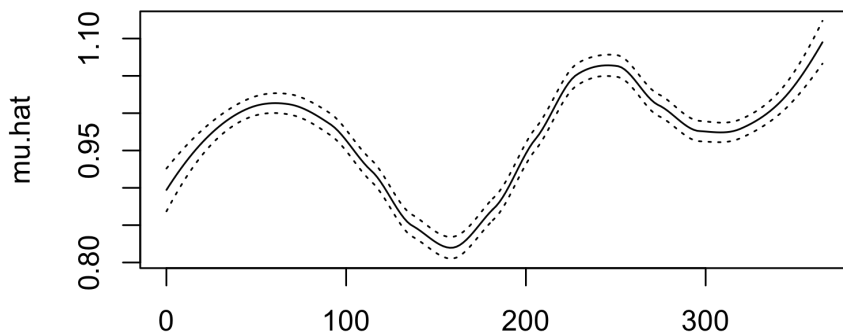


Figure 3: Estimate $\hat{\mu}_n$ of the mean hourly occupancy fraction μ_n . Time is shown in days.

B1. [2 points] Briefly describe what is a “local regression estimate”.

A window around each time point is used to construct a regression estimate. Here, linear regression is carried out over all the data points falling into the window.

B2. [2 points] The dashed lines in Fig. 3 show an approximate 95% confidence interval, constructed by adding $\pm 2SE$ where SE is the standard error on the estimate of the mean, as calculated by the local regression. Is this interval appropriate? Explain. Hint: it may help you to think about what you know about ordinary linear regression.

We know from Fig. 2A that the time series has considerable autocorrelation. An ordinary regression estimate still gives a reasonable point estimate in this situation, but the resulting standard errors should not be trusted.

B3. [2 points] Could the data be consistent with a model where the mean is not varying with time, e.g. a stationary process? Say yes or no, and explain.

Yes. The low frequency behaviour could be part of a random, long-term pattern.

SECTION C. We investigate whether a stationary model is appropriate for the detrended occupancy fraction $z_n = y_n - \hat{\mu}_n$. In particular, we compare the two time intervals August/September 2005 and March/April 2006. First, we fit an $ARIMA(1,0,1) \times (1,0,1)_{24}$ model to the 61 days in August and September 2005. Below is the R output.

```
##
## Call:
## arima(x = z[AugSep], order = c(1, 0, 1), seasonal = list(order = c(1, 0, 1),
## period = 24))
##
## Coefficients:
##          ar1          ma1          sar1          sma1  intercept
##          0.9139  0.0403  0.9998 -0.9884   -0.0060
## s.e.    0.0114  0.0277  0.0002  0.0080    0.1354
##
## sigma^2 estimated as 0.006561:  log likelihood = 1568.48,  aic = -3124.96
```

C1. [3 points] Write out the fitted model corresponding to the R output above, carefully stating all the model assumptions.

The fitted model is

$$(1 - 0.9990B^{24})(1 - 0.9139B)(y_t + 0.006) = (1 - 0.9884B^{24})(1 + 0.0403B)w_t,$$

where w_t are independent Gaussian random variables with mean 0 and variance 0.006561.

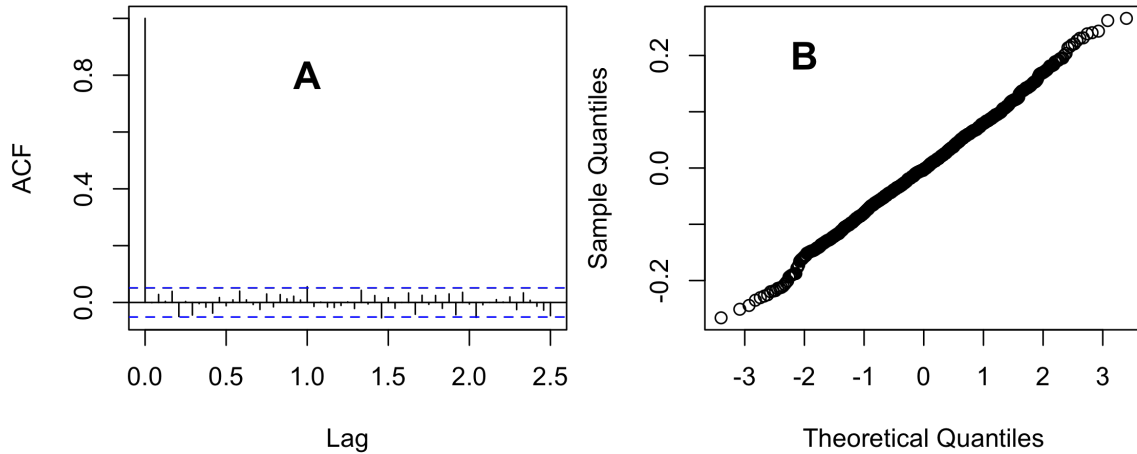


Figure 4: Investigation of the residuals from fitting an $ARIMA(1,0,1) \times (1,0,1)_{24}$ model to detrended occupancy for August/September. (A) Sample ACF. (B) Normal quantile plot, which plots the sorted residuals against the corresponding quantiles of the standard normal distribution.

C2. [2 points] What do you conclude from the diagnostic plots in Fig. 4? Also, explain at least one relevant property that is NOT checked by these diagnostic plots, and describe how you could check it.

From the normal quantile plot in B, we conclude that the distribution of the residuals is close to normal, which supports the normality assumption about w_t . From A, we conclude that the residuals appear to be uncorrelated. None of these diagnostics check for the stationarity of the process. One could check this by looking at the time plot of the residuals to see if there is any trend or pattern left and to see if the variability around the mean appears to be constant. Autocovariance stationarity could also be checked showing that the sample autocovariance function was similar at different intervals of the series.

C3. [3 points] Explain why the results in Fig. 4 and the R model for the fitted $ARIMA(1, 0, 1) \times (1, 0, 1)_{24}$ might support the choice of any of (a) $SARIMA(1, 0, 1) \times (1, 0, 1)_{24}$, (b) $SARIMA(1, 0, 1) \times (0, 1, 1)_{24}$, or (c) $SARIMA(1, 0, 1) \times (0, 0, 0)_{24}$. Explain which of these choices you think is best supported.

Fig. 4 is consistent with the choice of $ARIMA(1, 0, 1) \times (1, 0, 1)$. The estimate for the seasonal AR1 parameter, Φ_1 from the R output is very close to 1. When $\Phi_1 = 1$, an $ARIMA(1, 0, 1) \times (1, 0, 1)_{24}$ is equal to an $ARIMA(1, 0, 1) \times (0, 1, 1)_{24}$, so this should be an acceptable (and slightly simpler) model as well. However, on closer inspection, the two roots in the seasonal component almost cancel, which suggests instead considering $SARIMA(1, 0, 1) \times (0, 0, 0)_{24}$.

	MA0	MA1	MA2		MA0	MA1	MA2
(A) AR0	-378.9	-1612.9	-2258.7	(B) AR0	193.6	-1168.5	-1844.2
AR1	-3060.0	-3058.8	-3057.2	AR1	-2944.9	-2944.4	-2943.2
AR2	-3058.8	-3057.1	-3055.2	AR2	-2944.5	-2943.1	-2941.3
AR3	-3057.2	-3054.8	-3059.2	AR3	-2943.2	-2941.3	-2939.6

Table 1: AIC values from fitting $ARIMA(p, 0, q) \times (0, 1, 1)_{24}$ models to (A) August/September 2005, (B) March/April 2006.

SECTION D. We do some more analysis comparing the two time intervals August/September 2005 and March/April 2006.

D1. [2 points] A comparison of various models is presented in Table 1. Is there any conclusive evidence of imperfect likelihood maximization from these AIC values? Explain.

For example, -3054.8 for $ARMA(3, 1)$ is more than 2 higher than the $ARMA(2, 1)$ value of -3057.1. The true maximized log likelihood cannot be lower for $ARMA(3, 1)$ so the true AIC cannot be more than 2 higher.

D2. [2 points] What do you learn from the AIC values in Table 1 about choice of models for these data and the appropriateness (or otherwise) of fitting a stationary model to the entire time series.

We conclude from tables that an $ARIMA(1, 0, 0) \times (0, 1, 1)_{24}$ seems to be preferable, with the smallest AIC in both periods. As an extra word of caution, we cannot compare the AIC values in the tables with the -3124.96 value for $ARIMA(1, 0, 0) \times (1, 0, 1)$ because the data have been transformed, i.e. differences have been taken.

Below is the R output from fitting an $ARIMA(1, 0, 0) \times (0, 1, 1)_{24}$ model to detrended occupancy for August/September 2005 and March/April 2006.

```
##
## Call:
## arima(x = z[AugSep], order = c(1, 0, 0), seasonal = list(order = c(0, 1, 1),
##      period = 24))
##
```

```
## Coefficients:
##          ar1      sma1
##          0.9195  -1.0000
## s.e.    0.0104   0.0197
##
## sigma^2 estimated as 0.006496:  log likelihood = 1533,  aic = -3060
```

```
##
## Call:
## arima(x = z[MarApr], order = c(1, 0, 0), seasonal = list(order = c(0, 1, 1),
##      period = 24))
##
## Coefficients:
##          ar1      sma1
##          0.9436  -1.0000
## s.e.    0.0088   0.0341
##
## sigma^2 estimated as 0.007036:  log likelihood = 1475.46,  aic = -2944.92
```

D3. [4 points] Show how to use this output to carry out an approximate hypothesis test that the AR1 component is the same for August/September 2005 and March/April 2006 in the context of an $ARIMA(1, 0, 0) \times (0, 1, 1)_{24}$ model for detrended occupancy. Explain what your approximations are for this test. How good do you think these approximations are, and how could you check? Note: since you are not provided with statistical tables, you are not required to calculate a p-value.

Letting ϕ_1 and ϕ_2 be the AR1 coefficients for the first and second intervals, we wish to test the hypothesis

$$\begin{aligned} H_0 : \phi_1 &= \phi_2 \\ H_1 : \phi_1 &\neq \phi_2. \end{aligned}$$

Since we know that, asymptotically, $\hat{\phi}_i \sim N(\phi_i, \sigma_{\phi_i}^2)$, a test statistic that can be used (and is available from the R output above) is

$$Z = \frac{\hat{\phi}_1 - \hat{\phi}_2}{\sqrt{\hat{\sigma}_{\phi_1}^2 + \hat{\sigma}_{\phi_2}^2}},$$

where $\hat{\sigma}_{\phi_i} = \text{s.e.}(\hat{\phi}_i)$. Here, we would have to rely on the central limit theorem for the MLE and on the MLEs being uncorrelated for the 2 time intervals. The statistic comes out to be

$$z = \frac{0.9195 - 0.9436}{\sqrt{0.0104^2 + 0.0088^2}} = -1.769$$

This is within the acceptance region ± 1.96 , so we cannot reject the null hypothesis at the 5% level. The test seems to be based in enough data for the asymptotic approximation to be useful. One could check this and the assumed correlation by simulating from the fitted model and checking what the actual p-value is for some confidence level.