

# Time Series Analysis of Driver Performance

Megan Foster  
Sonali Nanavati  
Christina Webb

with the assistance of  
Brigeth Brookins  
Paul Sobanski

April 29, 2004

**Abstract:** Using data gathered by Johnson Controls, Inc., we examined the effect of the use of in-vehicle electronic devices on driving by means of a time series analysis. We determined which aspects of driver performance are significantly impacted by the use of these devices. We were then able to quantify the effect of driver distraction on measurable aspects of driver performance (eg. lane position, velocity) during the operation of in-vehicle electronic devices. The processes entailed in conducting a time series analysis are then generalized by a step-by-step procedure, providing guidance for further studies.

Work done for Johnson Controls, Inc. under the direction of C. Serafin and J. Kleiss, in partial fulfillment of the requirements of Michigan State University MTH 890, advised by Professor M. Winter.

# Table of Contents

Table of Contents.....	2
Introduction.....	3
Data Description.....	3
Time Series Analysis .....	3
Data Analysis .....	4
Process Model.....	9
Conclusion .....	10
Recommendations .....	10
Appendix A. Data .....	12
Appendix B. Autocorrelation Program.....	13
Appendix C. STATA Output .....	14
Appendix D. Process Model.....	17

## Introduction

Driver distraction is a topic of growing concern with the increasing use of in-vehicle devices. We investigated the degree to which a driver's performance is affected by operating electronic devices. Most driver performance analyses examine this relationship over entire time intervals. Time data is grouped together and then averages are computed and compared. However, driving is a continuous process, so it was our goal to develop a process model that can be used for continuous analysis.

Using our process model, we studied driver performance data obtained from a driving simulator. The effects of operating devices, such as radios and cellular telephones, were considered. From the data provided, we examined the continuing changes in driver performance due to the operation of these devices.

## Data Description

A common method for assessing driver distraction is to measure a number of driver performance variables in a driving simulator. A driver navigates a simulated course and is instructed to execute different tasks. These tasks include the operation of a cellular phone and radio.

During the simulation, several indices are recorded every tenth of a second to measure driver performance. The variables considered are velocity, lane position, steering angle, acceleration, braking frequency, distance from a leading car, and hand position while driving. A marker is inserted in the data when the driver is engaged in a task.

Previously, similar data has been analyzed as an average across a group of drivers in discrete time intervals [1]. Thus, the continuous data was lumped together to form discrete data, and then analyzed. This analysis fails to capture the continuous effects of the use of in-vehicle devices on driving performance. Yet, the subtle changes in the continuous variables yield valuable insight into modeling this process. Thus, to study the data, a time series analysis is needed.

## Time Series Analysis

A time series is a collection of observations made sequentially in time. The first step in the analysis of a time series is to produce a description of the data by graphing a measured variable against time. This allows us to see possible sources of variation, such as the beginning of a task. At this point in the analysis, outliers may be identified. To keep the analysis accurate, these outliers may need to be adjusted [2].

A key characteristic of time series analysis is the autocorrelation coefficient,  $r_k$ . These coefficients measure the correlation between observations of a variable with itself lagged a distance apart. This may provide insight into the probability model, which generated the data. The autocorrelation coefficients are defined by

$$r_k = \frac{\sum_{t=1}^{N-k} (x_t - \bar{x})(x_{t+k} - \bar{x})}{\sum_{t=1}^N (x_t - \bar{x})^2}, \quad (1)$$

where  $k$  is the lag, or the  $k^{\text{th}}$  data point behind the original.

A useful aid in interpreting a set of autocorrelation coefficients is a graph called a *correlogram*, in which  $r_k$  is plotted against the lag  $k$ . The shape of the correlogram is indicative of the appropriate model for a given time series. If the graph approaches zero within the first few lags, then the data is considered *stationary*. This means the properties of the underlying model do not change through time and that there is no trend in the data [2]. If the correlogram does not approach zero quickly, the series must be differenced to ensure the validity of the regression analysis. During this process, a new series,  $\{?x_t\}$  is created where

$$\Delta x_t = x_t - x_{t-1}. \quad (2)$$

We then compute the autocorrelation coefficients of  $\{?x_t\}$  and graph the correlogram. If the correlogram still does not approach zero within the first few lag values, the differencing process is repeated until the data is considered stationary [3].

After ensuring a stationary process, smoothing may be necessary. This is an especially useful procedure for data in which emphasis on variation at particular times is important [2]. Statistical software packages have several smoothing functions designed for different types of data. In an event analysis, a locally weighted logarithmic smoothing method should be used [4].

The *autoregressive integrated moving average* (ARIMA) class of models is an important forecasting tool and is the basis of many fundamental ideas in time series analysis [5]. An ARIMA model has two components---an autoregressive (AR) and a moving average (MA) process. A time series,  $\{x_t\}$ , is considered an AR process of order  $p$  if it is a weighted linear sum of the previous  $p$  time values, plus a random component,  $Z_t$ , such that

$$x_t = \mathbf{f}_1 x_{t-1} + \mathbf{f}_2 x_{t-2} + \dots + \mathbf{f}_p x_{t-p} + Z_t. \quad (3)$$

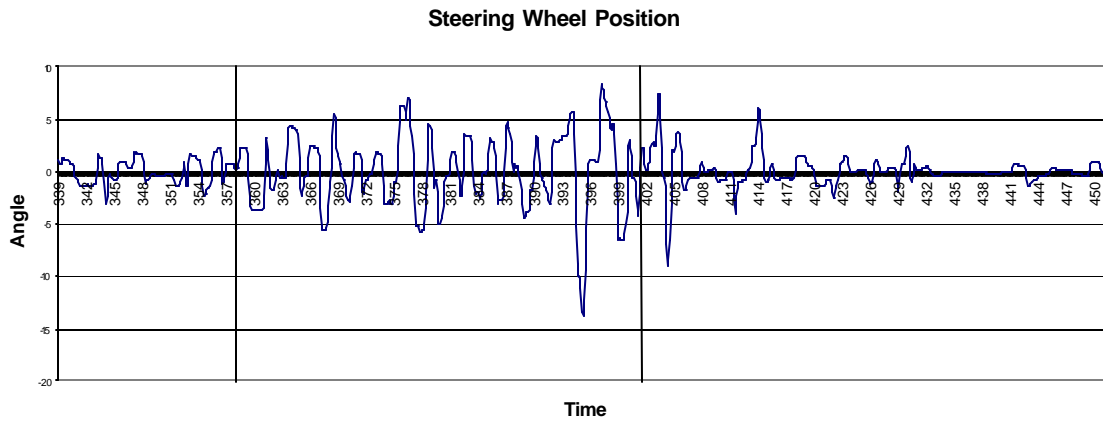
Here,  $Z_t$  denotes a purely random process with zero mean and constant variance. A time series is considered to be an MA process of order  $q$  if it is a weighted linear sum of the last  $q$  random components, such that

$$x_t = Z_t + \mathbf{q}_1 Z_{t-1} + \dots + \mathbf{q}_q Z_{t-q}. \quad (4)$$

In an event analysis, only the AR component is necessary for the model.

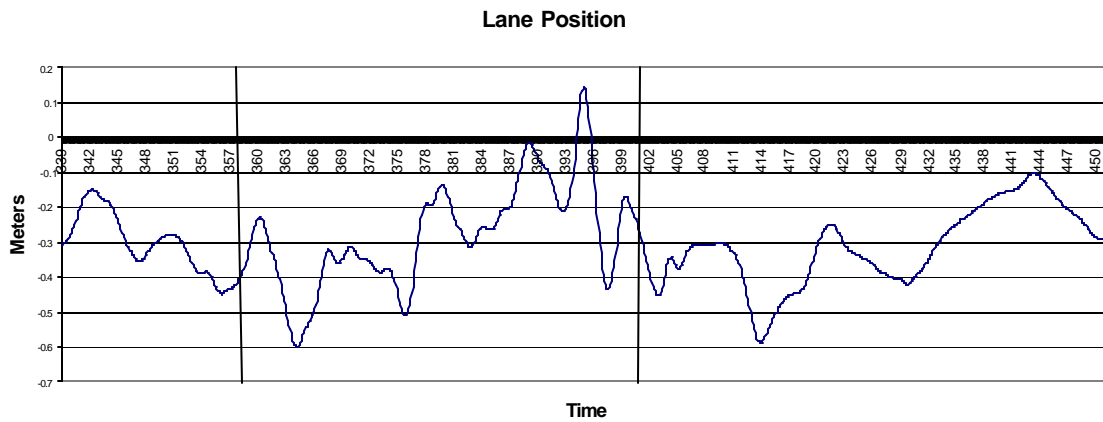
## Data Analysis

In order to conduct our analysis, graphs were generated to compare the response of each variable throughout the time interval. Figure 1 depicts the angle of the steering wheel while in the driving simulator. A negative wheel angle indicates veering to the left, while a positive angle shows a turn to the right. The vertical lines indicate the beginning and the end of a cell phone conversation. As can be seen in Figure 1, there are frequent and dramatic shifts in steering while the driver is engaged in conversation, even though the course required no turning. This shows the negative effects of cell phone use on driving.



**Figure 1.** Graph of the subject's steering wheel position versus time in the simulator. Vertical lines represent the onset and completion of the assigned task (cell phone conversation).

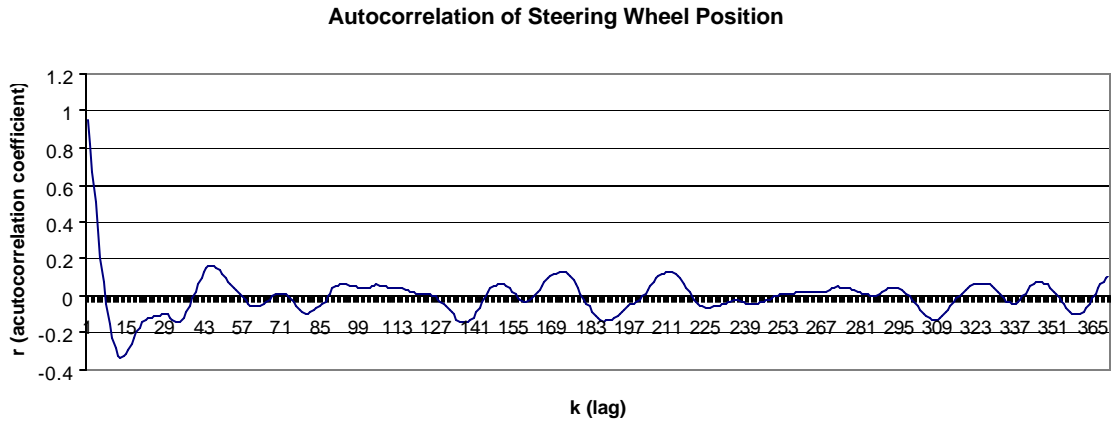
Figure 2 depicts the lane position of the test subject versus time. The lane position graph represents the subject's deviation from the center of the lane. The center is represented as zero. A deviation to the left is a negative value and to the right is positive value. As can be seen in Figure 2, the driver tends to maintain a negative lane position. This is considered a trend, and thus it is necessary to difference the data before developing a model. In addition, the onset of the task causes greater variation in the range of lane position values. Thus, the cell phone conversation negatively impacts the driver's ability to maintain a constant lane position [5].



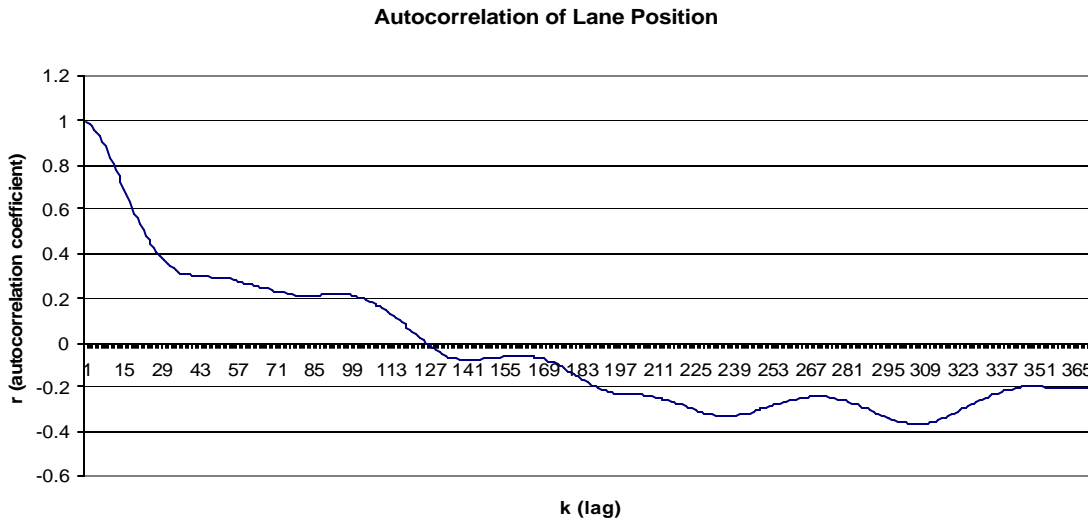
**Figure 2.** Graph of the subject's lane position versus time in the simulator. Vertical lines represent the onset and completion of the assigned task (cell phone conversation).

To quantify the driver distraction, we conducted a time series analysis. First, we computed the autocorrelation coefficients using Equation 1. A Matlab program was written (Appendix B) to compute these values efficiently. We then graphed the correlograms for each measured variable.

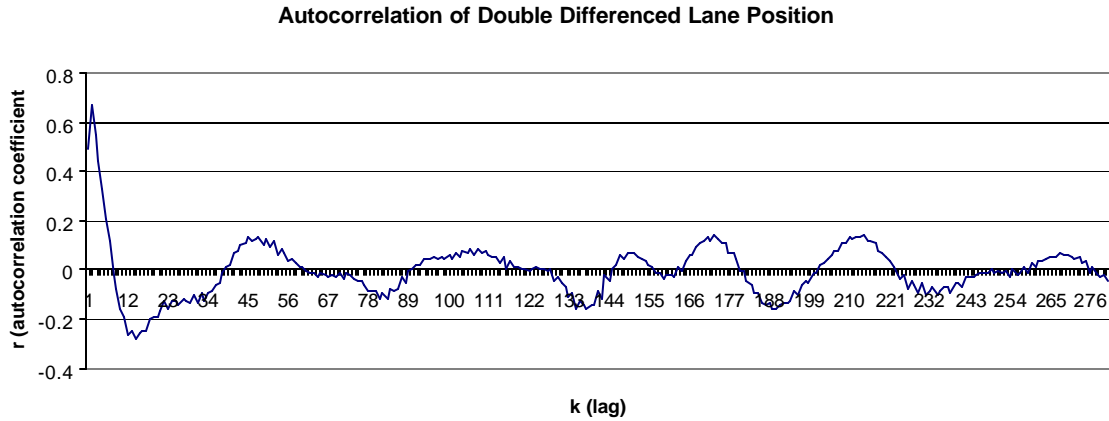
The autocorrelation coefficients for steering wheel position, as shown in Figure 3, decrease quickly to zero within the first few lags. This means the data is stationary and ready for regression analysis. However, the coefficients for lane position (Figure 4) take significantly longer to converge to zero. Because of this, we differenced the lane position data. After performing the differencing process twice, we produced the correlogram in Figure 5, which shows the data now approaching zero within the first few lag values.



**Figure 3.** This figure depicts the correlogram for steering wheel position. This correlogram approaches zero within the first few lag values, signaling the data is stationary.



**Figure 4.** This figure depicts the correlogram for lane position. The correlogram does not approach zero for small lag values, signaling the data is not stationary.



**Figure 5.** This figure depicts the correlogram for the double differenced lane position data. This correlogram approaches zero within the first few lag values, signaling the data is stationary.

The next step in the analysis was to smooth the data. To do this, we used a locally weighted logarithmic smoothing function in STATA, a statistical software package. The weights were computed based on whether or not the task was being executed.

After smoothing, the data was analyzed using an AR model. We chose a model of order one. This was because increasing the order did not significantly reduce the standard error. Since we were determining the effect of the task on driver performance, we created a binary variable, *event*, which in this example identifies the subject’s involvement in the task at time *t*. The structure of the model is

$$\hat{x}_t = \mathbf{a}_0 + \mathbf{a}_1 event_t + \mathbf{f}x_{t-1}, \tag{5}$$

where  $x_t$  denotes a stationary process, either originally or from differencing. The specific model equations are shown below in Table 1.

**Table 1.** This table summarizes the model equations found using STATA for steering angle, lane position, and velocity based on the simulation data from a cell phone conversation.

Steering Angle	$\hat{x}_t = 0.42 + 0.17event_t + 0.88x_{t-1}$	$p = 0.00$
Lane Position	$\hat{x}_t = 0.49 + 0.03event_t + 0.96x_{t-1}$	$p = 0.31$
Velocity	$\hat{x}_t = -7.02 - 0.84event_t + 0.02x_{t-1}$	$p = 0.95$

In the above equations, the *p* values indicate the significance level of the task. Generally, a *p* value of 0.10 or lower indicates the task has a significant effect. If the *p* value is above 0.10, the significance of the task cannot be determined. For steering angle,  $p = 0.00$ , which means that performing the task (a cell phone conversation) has an impact on the subject’s steering ability. On the other hand, the *p* value for velocity is 0.95 meaning the subject’s velocity is not significantly affected by performing the task.

The coefficient of the task,  $a_1$ , is the average effect of the task on the studied variable. The  $a_1$  value for steering angle is 0.17. This means that steering angle will deviate on average by 0.17 as a result of the task. This does not measure the direct effect of the task. An estimator is used to more explicitly determine the effect of the task [3]. For our general equation, this estimator is

$$\mathbf{w} = \frac{\mathbf{a}_1}{1 - \mathbf{f}}. \quad (6)$$

This measures the change in the variable values as a direct result of performing the task. A larger magnitude of  $\mathbf{w}$  means the effect of the task on the measured variable is more drastic. Specifically, for steering angle,  $\mathbf{w} = 1.44$ . Thus, steering angle changes by a factor of 1.44 while talking on a cell phone. This is the quantifiable effect of a cell phone conversation.

To study the effects of a cell phone conversation at certain points in time, the *event* variable can be modified. Specifically, to study the differences in lane position at the onset versus the remaining time engaged in conversation, we concentrated on the data while on the cell phone. To do this, we removed all data during the times when the subject is not engaged in the task. The binary variable *event* was changed to ones at the beginning of the task and zeros thereafter. Next, the ARIMA model was found. This model along with another model of the effect on lane position immediately following the task is shown in Table 2. For the end of task model, we concentrated on the interval of time immediately following the task (cell phone conversation), comparing it to the data thereafter.

**Table 2.** This table summarizes the model equations found using STATA for lane position after the cell phone conversation and lane position at the onset of the conversation.

Lane Position (onset of task)	$\hat{x}_t = -6.78 - 0.19event_t + 0.89x_{t-1}$	$p = 0.10$
Lane Position (end of task)	$\hat{x}_t = -7.29 + 0.41event_t + 0.94x_{t-1}$	$p = 0.06$

The  $p$  values for the onset and end of task effect on lane position are 0.10 and 0.06 respectively. This means that lane position is significantly impacted at the onset of the task, as well as immediately following the task. The quantifiable effect,  $\mathbf{w}$ , at the onset of the task on lane position is 1.73. This means the difference between the lane position values at the onset of the task versus the remaining task values changes by a factor of 1.73. The effect on lane position at the end of the task is 6.83, meaning that even after the task has ended the driver's performance is still significantly impacted as a result of the task. In Table 1, we found that the overall effect on lane position of a cell phone conversation is not significant ( $p = 0.31$ ). However, we can see that the initial and lingering effects of cell phone conversation are substantial.

## Process Model

We conducted our time series analysis on specific data. However, this process can be generalized to study any data set. This step-by-step process is easily represented by a flowchart, Appendix D. The following is a detailed description of each stage of the process:

**Step 1:** Create a graph of the data versus time.

This allows any obvious characteristics and trends to be identified.

**Step 2:** Compute the autocorrelation coefficients.

Using the MATLAB algorithm in Appendix B, the autocorrelation coefficients can be calculated from data read in from Excel files. The autocorrelation coefficients are outputted to a separate Excel file.

**Step 3:** Construct the correlogram.

Using the output from Step 2, graph the autocorrelation coefficients versus the lag.

*3a.* Determine if the graph quickly approaches zero (within the first few lag values) and then continues to oscillate around zero. If “yes”, move to Step 4, otherwise continue to 3b.

*3b.* Difference the data by subtracting two consecutive values to create a new variable for study. Return to Step 2.

**Step 4:** Isolate event to study.

Choose the time period of interest and remove all other time values. Create a binary *event* variable. Place ones during the event of interest and zeros elsewhere.

**Step 5:** Smooth the data.

Smooth the data using a built in locally weighted log smoothing function. The smoothing is weighted based on the *event* variable.

**Step 6:** Generate the ARIMA model.

Using statistical tool package, find the ARIMA model.

*6a.* Identify the  $p$  value of the *event* variable. If this  $p$  value is less than 0.1, continue to 6c. Otherwise, go to 6b.

*6b.* It cannot be determined that the event has a significant effect on the measured variable. Thus, this is the conclusion of the analysis for this event.

*6c.* The event has a significant effect on the measured variable.

**Step 7:** Calculate and interpret effect.

Compute the effect estimator,  $w$ . A large magnitude of  $w$  indicates a greater effect of the event on the measured variable.

## Conclusion

The continuous effect of the use of in-vehicle electronic devices on driver performance is quantified using a time series analysis. The process model describes each step of this process so it can be easily applied to any set of time data. By changing the *event* variable, any time period of interest can be studied and its continuous effect quantified. Specifically, we found that an individual's steering ability is affected by engaging in a cell phone conversation, as well as his lane position at the beginning of, and just after a conversation.

## Recommendations

In this final section, we indicate several directions of research for consideration in the future.

### **Recommendation 1: Examine long-term changes.**

We recommend performing a time series analysis on the long-term changes in driver distraction. For example, we have shown cell phone use initially has a great impact on an individual's driving. However, over time an individual becomes more familiar with driving while engaged in conversation. Thus, the estimated effect of a cell phone conversation is high with the first use, lower with the tenth use, even lower with the 100<sup>th</sup> use, and so on. By performing a time series analysis on driver performance, this diminishing effect can be quantified and the long-term effect can be measured. Here, the time variable will be the number of times the subject has driven while engaged in conversation.

### **Recommendation 2: Consider controlled versus simulated driving.**

We also feel it would be beneficial to analyze data gathered from controlled driving experiments on actual roads, rather than from a simulator. There are no real consequences in a simulator for decreased driver performance, thus drivers may tend to be more careless than in a real world situation. Alternatively, a simulated path should be created that models actual driving situations by incorporating other vehicles, signs, pedestrians, etc.

### **Recommendation 3: Compare city versus highway driving situations.**

Lastly, we recommend comparing the effects of in-vehicle devices on city versus highway driving. The demands on drivers in both of these situations are unique. Due to this, the effects of in-vehicle devices in these two situations may be substantially different.

## References

- [1] Kleiss, James, *Interview: Johnson Controls Inc*, Feb. 19, 2004.
- [2] C. Chatfield, *The Analysis of Time Series: An Introduction*, New York: Chapman and Hall, 1989, pp. 1-25.
- [3] J. Wooldridge, *Introductory Econometrics: A Modern Approach*, Ohio: Thompson Learning, 2003, pp. 323-421.
- [4] *STATA Reference: Release 6*, Texas: Stata Press, 1999, pp 152 – 155.
- [5] C. Chatfield, *Time-Series Forecasting*, New York: Chapman and Hall, 2001, pp. 1-30.

## Appendix A. Data

**Table 3.** This table is an excerpt of the data given by Johnson Controls, Inc. Each column represents a measured variable during part of a particular driving simulation. The digital inputs column indicates the beginning of a task when the column changes from 9 digits to 12 digits. The ones and zeros represent the number of hands the driver has on the wheel during the simulation.

**Time { Nov 14, 2003 2:09:37 PM }**

Time	Velocity	LanePos	Steer	Accel	Brake	HeadwayDist	DigitalInputs	LeadVelocity
173.7201	14.852	-0.24	1.3	0.097	0	59.09	101001001	15.641
173.8201	14.865	-0.241	2.1	0.105	0	59.168	101001001	15.641
173.9202	14.88	-0.24	2	0.114	0	59.244	101001001	15.641
174.0201	14.898	-0.24	1.2	0.122	0	59.32	101001001	15.641
174.1202	14.919	-0.238	0.1	0.137	0	59.394	101001001	15.641
174.2202	14.951	-0.237	-0.8	0.19	0	59.465	101001001	15.641
174.3202	15.005	-0.236	-1.1	0.193	0	59.531	101001001	15.641
174.4202	15.06	-0.235	-1.3	0.193	0	59.592	101001001	15.641
174.5202	15.113	-0.234	-1.4	0.193	0	59.648	101001001	15.641
174.6202	15.166	-0.233	-1.5	0.193	0	59.698	100001001	15.641
174.7202	15.219	-0.234	-1.5	0.193	0	59.743	100001001	15.641
174.8202	15.271	-0.235	-1.8	0.193	0	59.783	100001001	15.641
174.9202	15.319	-0.237	-2.7	0.126	0	59.818	100001001	15.641
175.0202	15.34	-0.24	-3.3	0.119	0	59.85	100001001	15.641
175.1202	15.358	-0.244	-3	0.131	0	59.88	100001001	15.641
175.2202	15.385	-0.249	-2.1	0.169	0	59.908	100110001001	15.641
175.3202	15.427	-0.255	0	0.183	0	59.932	100110001001	15.641
175.4202	15.476	-0.261	2.7	0.185	0	59.951	100110001001	15.641
175.5202	15.524	-0.267	4.5	0.185	0	59.965	100110001001	15.641
175.6202	15.571	-0.271	5.2	0.185	0	59.974	100110001001	15.641
175.7202	15.617	-0.274	5.2	0.185	0	59.979	100110001001	15.641
175.8202	15.663	-0.275	5.1	0.185	0	59.979	100110001001	15.641
175.9202	15.709	-0.274	5.1	0.185	0	59.975	100110001001	15.641
176.0202	15.756	-0.27	5.1	0.185	0	59.967	100110001001	15.641
176.1202	15.802	-0.264	4.9	0.185	0	59.954	100110001001	15.641
176.2202	15.848	-0.255	4.7	0.184	0	59.936	100110001001	15.641
176.3202	15.885	-0.245	4.6	0.061	0	59.913	100110001001	15.641
176.4202	15.862	-0.232	4.2	0.002	0	59.89	100110001001	15.641
176.5202	15.821	-0.217	3.8	0.002	0	59.871	100110001001	15.641
176.6202	15.785	-0.2	3.1	0.002	0	59.857	100110001001	15.641
176.7202	15.751	-0.182	2.1	0.002	0	59.845	100110001001	15.641
176.8202	15.717	-0.163	0.5	0.002	0	59.836	100110001001	15.641
176.9202	15.683	-0.143	-1.6	0.002	0	59.831	100110001001	15.641
177.0202	15.649	-0.124	-2.7	0.002	0	59.828	100110001001	15.641
177.1202	15.613	-0.105	-2.7	0.002	0	59.829	100110001001	15.641
177.2202	15.578	-0.087	-2.7	0.002	0	59.834	100110001001	15.641

## Appendix B. Autocorrelation Program

**Table 4.** The Matlab algorithm computes the autocorrelation coefficients for lag values ranging from one to approximately one fourth the number of data points. The autocorrelation coefficients are used to determine the nature of the data. Data is imported from an Excel workbook and the autocorrelation coefficients are exported to another Excel file for easy interpretation.

```
M = wklread('velocity.wkl')           % reads in the renamed Excel file
                                       % for a specified variable & task

N=size(M);                             % number of points
k = round(N/4);                         % lag sizes

xbar=mean(M);                           % computing the mean
numerator=0;                            % initializing autocorrelation
r=zeros(k,1);                           % vector

for i=1:N
    numerator=numerator + (M(i)-
xbar)^2;
end;

                                       % computes the denominator of the
                                       % autocorrelation function
for i=1:k
    denominator=0;
    for j=1:N-i
        denominator = denominator +
(M(j)- xbar)*(M(j+i)-xbar);
    end;
    r(i)=numerator/denominator;         % stores value in the vector
end;

                                       % writes the data to an Excel file
wklwrite('velocityoutput.wkl', r);    % for graphing
```

## Appendix C. STATA Output

**Table 5.** This table contains the ARIMA analysis by STATA for the smoothed steering angle data. The data was collected before, during, and after engaging in a cell phone conversation. The driver's performance during the task was compared to his performance surrounding the task.

### Smoothed Steering

ARIMA regression

```
Sample: 1 to 1126                      Number of obs   =    1126
                                         Wald chi2(2)    =   2720.33
Log likelihood = 1173.209                Prob > chi2     =    0.0000
```

		Coef.	OPG Std. Err.	z	P> z	[95% Conf. Interval]	
s1							
event		.1743088	.0376309	4.632	0.000	.1005536	.248064
_cons		.4235896	.040488	10.462	0.000	.3442344	.5029447
ARMA							
ar							
	L1	.8792677	.0169871	51.761	0.000	.8459736	.9125617
/sigma		.085305	.0010707	79.670	0.000	.0832064	.0874035

**Table 6.** This table contains the ARIMA analysis by STATA for the smoothed double differenced lane position data. The data was collected before, during, and after engaging in a cell phone conversation. The driver's performance during the task was compared to his performance surrounding the task.

### Smoothed Double Differenced Lane Position

ARIMA regression

```
Sample: 2 to 1126                      Number of obs   =    1125
                                         Wald chi2(2)    =  12653.50
Log likelihood = 1854.55                Prob > chi2     =    0.0000
```

		Coef.	OPG Std. Err.	z	P> z	[95% Conf. Interval]	
l1							
event		.0318161	.0313164	1.016	0.310	-.0295628	.0931951
_cons		.4901455	.0474278	10.335	0.000	.3971887	.5831024
ARMA							
ar							
	L1	.9633929	.0086015	112.003	0.000	.9465343	.9802515
/sigma		.0464863	.000576	80.711	0.000	.0453574	.0476151



**Table 9.** This table contains the ARIMA analysis by STATA for the smoothed differenced end of task lane position data. The data was compared immediately following and with data unaffected by engaging in a cell phone conversation.

**Smoothed Differenced End of Task Lane Position**

ARIMA regression

```

Sample: 616 to 1126                Number of obs   =      511
                                   Wald chi2(2)      =     180.59
Log likelihood = -91.59064          Prob > chi2     =      0.0000
    
```

```

-----+-----
dlpsm |          Coef.      OPG      Std. Err.      z    P>|z|      [95% Conf. Interval]
-----+-----
dlpsm |
event |      .4081946      .2124634      1.921    0.055      -.008226      .8246152
_cons |     -7.291009      .565949     -12.883    0.000      -8.400249     -6.181769
-----+-----
ARMA
ar
  L1 |      .9413095      .0725014     12.983    0.000      .7992094      1.08341
-----+-----
/sigma |      .2888563      .0110304     26.187    0.000      .2672371      .3104756
-----+-----
    
```

## Appendix D. Process Model

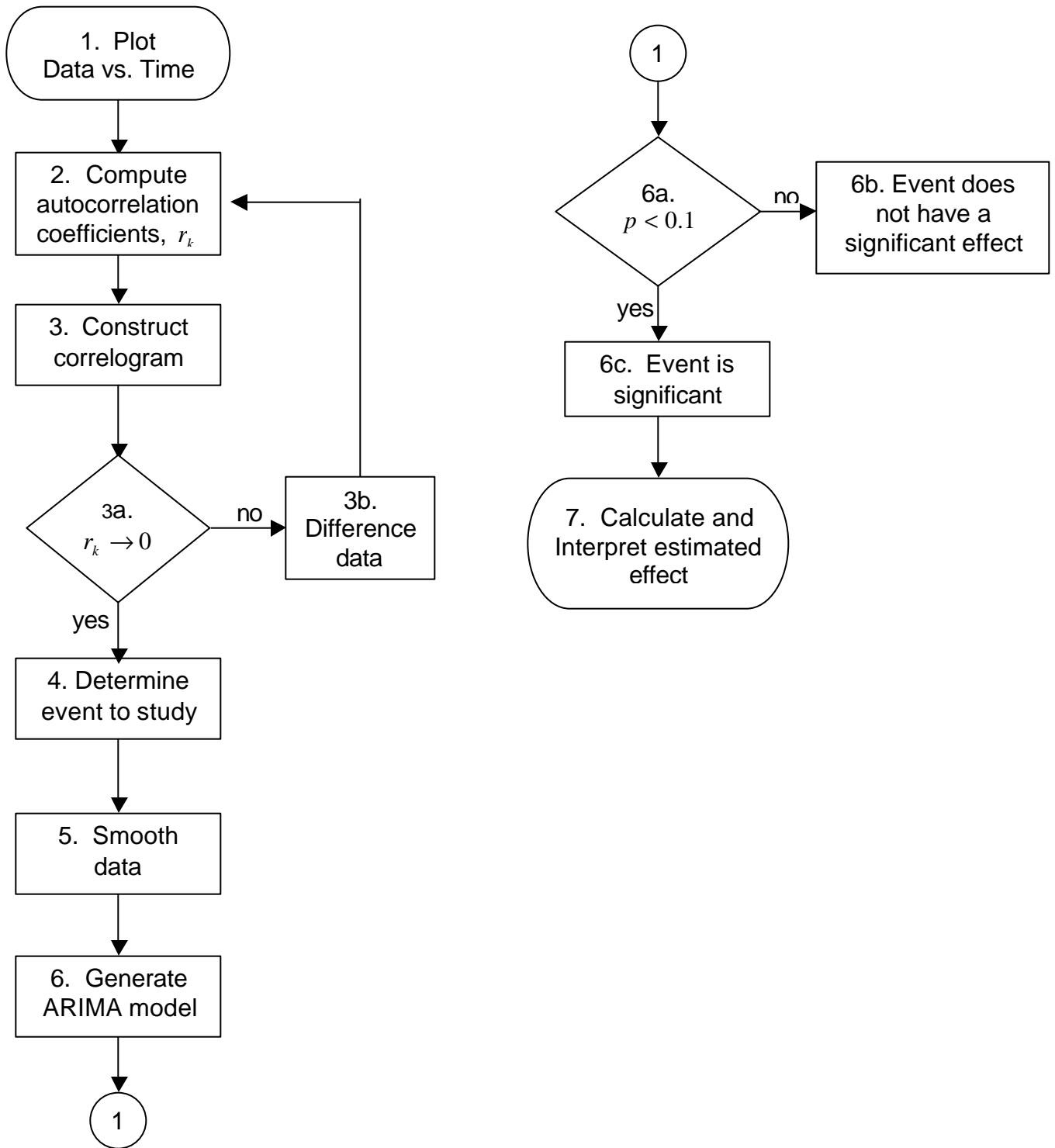


Figure 6. The flowchart models the step-by-step process of a time series analysis of driver distraction.