Objectives

Introduce the SPSS define dates option.
Run a sequence chart in SPSS
Explain how to deal with missing time series data.

Starting Time Series Analysis

Data

The data used in this section are from a private mailing company, measuring the volume of mail delivered each day of the week, including weekends, over a four-week period. The data file has two missing values where data had not been collected on the number of parcel deliveries for those days.

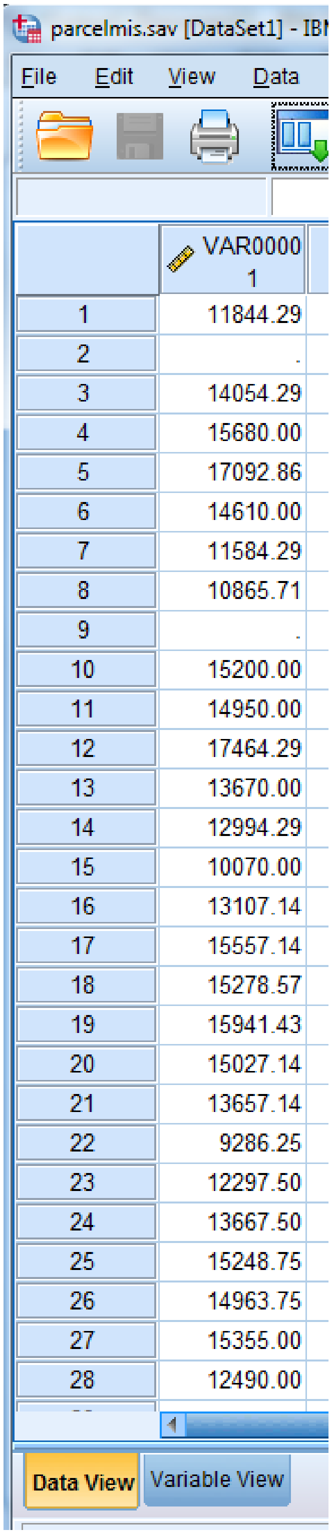
Defining Dates and Displaying Time Series Data in SPSS

Prior to undertaking any time series analysis in SPSS, it is always important to set up time series dates using the Define Dates dialog from the Data menu. If dates are not set up using the Define Dates procedure, SPSS will often assume that the data are not suitable for time series. As a result, it will not be possible to run certain time series models.

It is not, for example, sufficient to define date variables in a spreadsheet, such as Microsoft Excel, and then import the date variables into SPSS. While the date variables will be converted into SPSS as dates, the overall data set that is created will not be recognized as a data set suitable for time series analysis and the data periodicity will also not be defined. It is through the **Define Dates** procedure that SPSS will be given information on the periodicity of the data.

Create a file called parcelmis.sav and input the following data:

Figure 2.2 Parcel Data File



	VAR00001
1	11844.29
2	.
3	14054.29
4	15680.00
5	17092.86
6	14610.00
7	11584.29
8	10865.71
9	.
10	15200.00
11	14950.00
12	17464.29
13	13670.00
14	12994.29
15	10070.00
16	13107.14
17	15557.14
18	15278.57
19	15941.43
20	15027.14
21	13657.14
22	9286.25
23	12297.50
24	13667.50
25	15248.75
26	14963.75
27	15355.00
28	12490.00

To assign the appropriate dates to each row in the data editor:

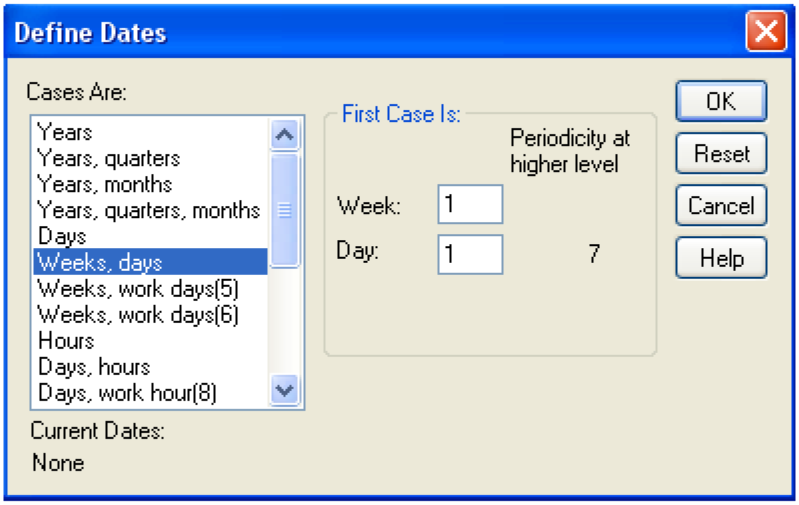
Click **Data...Define date and time**

When the Define Dates dialog box opens it shows the data status of the file in the lower-left corner under Current Dates We see that the file is not yet dated because the dialog box reads "*Current Dates: None*." In the *Cases Are* list box there are options allowing for a wide range of date specifications, including yearly, monthly, quarterly, daily and hourly data. The data collected in the *Parcelmis* file are daily data for a parcel company which offered a seven-day parcel delivery service. We therefore need to set up daily time series dates.

Select **Weeks, days** in the Cases Are list box
In the **First Case Is** text box for **Week** enter **1**
Enter **1** in the **Day** text box

Notice that the Define Dates dialog shows the periodicity of the daily time series; it is set to seven for the Weeks, Days choice. Also, the second specification for Day (7) sets the period: seven days form the period.

Figure 2.3 Defining Time Series Dates in SPSS



First Case Is:

Periodicity at higher level

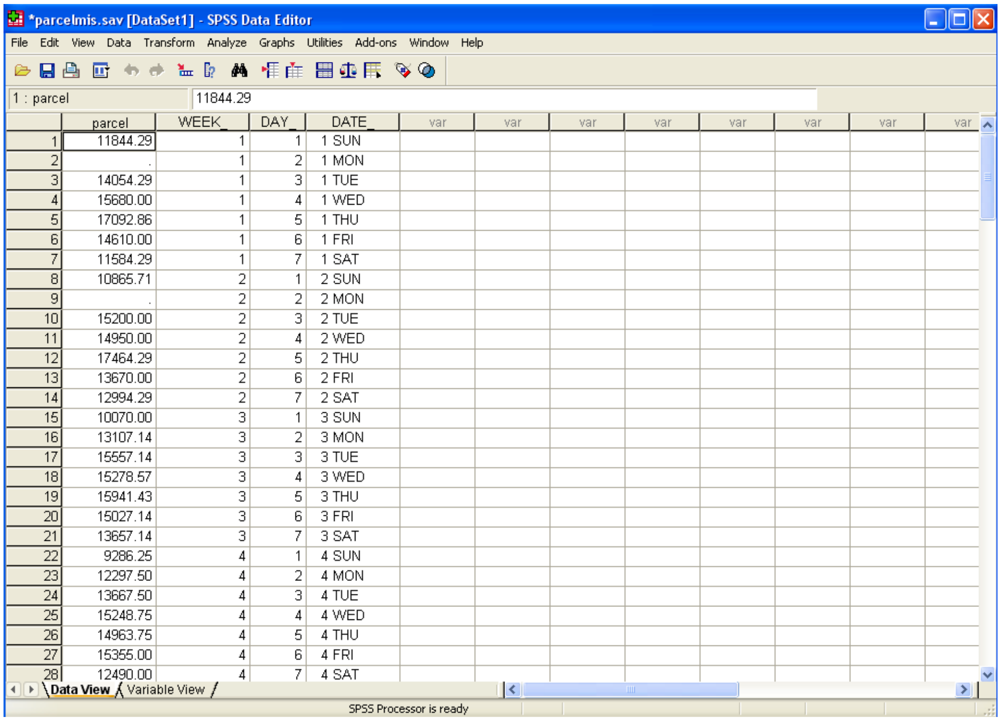
Week: 1

Day: 1 7

Current Dates:
None

Click on **OK** to process the request
Move to the **Data Editor** window

Figure 2.4 Newly Created Date Variables


*parcelmis.sav [DataSet1] - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help
Data View
Variable View

1 : parcel 11844.29

	parcel	WEEK_	DAY_	DATE_	var	var	var	var	var	var	var	var
1	11844.29	1	1	1 SUN								
2	.	1	2	1 MON								
3	14054.29	1	3	1 TUE								
4	15680.00	1	4	1 WED								
5	17092.86	1	5	1 THU								
6	14610.00	1	6	1 FRI								
7	11584.29	1	7	1 SAT								
8	10865.71	2	1	2 SUN								
9	.	2	2	2 MON								
10	15200.00	2	3	2 TUE								
11	14950.00	2	4	2 WED								
12	17464.29	2	5	2 THU								
13	13670.00	2	6	2 FRI								
14	12994.29	2	7	2 SAT								
15	10070.00	3	1	3 SUN								
16	13107.14	3	2	3 MON								
17	15557.14	3	3	3 TUE								
18	15278.57	3	4	3 WED								
19	15941.43	3	5	3 THU								
20	15027.14	3	6	3 FRI								
21	13657.14	3	7	3 SAT								
22	9286.25	4	1	4 SUN								
23	12297.50	4	2	4 MON								
24	13667.50	4	3	4 TUE								
25	15248.75	4	4	4 WED								
26	14963.75	4	5	4 THU								
27	15355.00	4	6	4 FRI								
28	12490.00	4	7	4 SAT								

SPSS Processor is ready

Three new variables have been created in the Data Editor window. The *WEEK_* and *DAY_* variables are numeric, while *DATE_* is a string variable.

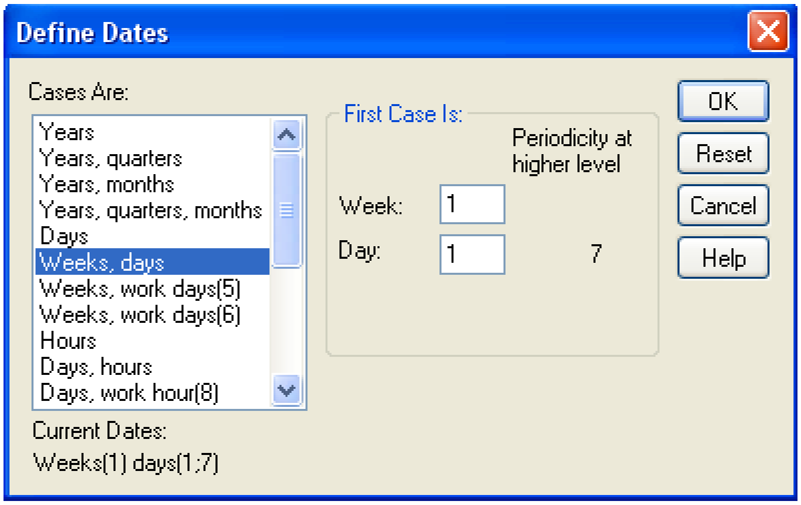
Note

The Define Dates procedure simply assigns dates to each case in succession based on the type of dating scheme selected in the *Cases Are* list box and the first case specifications entered in the *First Case Is* area. Thus, you must ensure that your data file contains a case for every intermediate time period between the first and the last. Even if you do not have a data measurement for a particular time period, you still must have a case in the file for that time period. The file shown in Figure 2.4 is a good example of this. We see that two Monday parcel measurements are missing, but the analyst placed empty cases in the file nevertheless. It is necessary to do this so the subsequent data values will apply to the correct time points. All the SPSS time series analysis procedures assume that all time periods physically exist as cases in the file and will not operate correctly unless they do.

Returning to the Define Dates dialog box shows that the data are now being treated as daily time series data. If dates are not defined in this way, all the time series dialog boxes will show the current periodicity as none.

Click **Data...Define date and time** from the Data Editor window

Figure 2.5 Time Series Dates Set-Up by the Define Dates Dialog



Week: 1

Day: 1 7

Current Dates:

Weeks(1) days(1;7)

This confirms that SPSS now recognizes the data as being a daily time series with a periodicity of seven, as shown in the Current Dates indicator in the lower left corner of the dialog box.

Dealing With Missing Values

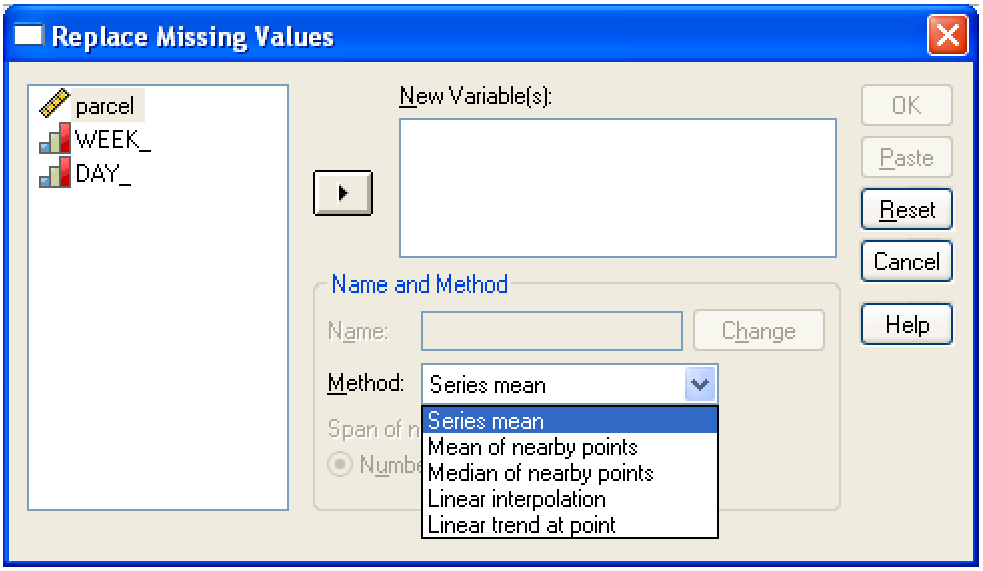
Data collection is often fraught with difficulties and data may not be recorded for every time period of interest. Missing data at the ends of a time series pose no particular problem apart from shortening the length of the series. However, some time series routines will not function if there are missing data at points other than at the beginning or end of the series (in other words, even if there are cases defined for those dates, the lack of a valid value causes difficulty). Missing data embedded in a time series can therefore be a serious problem for certain time series routines. In order to overcome this problem it is often necessary to "splice" the series. Basically, splicing is a method used to replace missing data with numeric information from elsewhere in the time series.

In order to overcome the problem of missing values, SPSS has an option called **Replace Missing Values** which is specifically designed for splicing. Looking in the Data Editor, it is apparent that for the first two weeks, data were not collected on Mondays. In this instance it is therefore necessary to replace the missing values with information collected on other dates. Unlike in many standard statistical analyses (e.g., with survey data), estimating missing data is common in time series analysis.

From the Data Editor window:

Click **Transform…Replace Missing Values**
Click **Method** drop-down list

Figure 2.6 Replace Missing Values Dialog Box



The numeric variables in the Data Editor are displayed in the variables list box of the Replace Missing Values dialog box.

Within this dialog box there are many options for replacing missing values with combinations of non-missing values. The *Method* drop-down list contains options to replace missing values using one of several time-series functions.

- The first option, *Series mean*, replaces missing values in the series with the overall mean of the series.

- For the *Mean of nearby points* method, one of the two *Span of nearby points* options must be selected. If the *Number* option is selected, the specified number of valid values above and below the missing data points will be used to compute a mean value to replace the missing value (the default is 2). Alternatively, the *All* option can be specified, and this is equivalent to the series-mean option
- The *Median of nearby points* is the same as the *Mean of nearby points* option, except the median is used to replace the missing value.

- The *Linear interpolation* option is useful for replacing missing values when there is more than one consecutive missing value in the series. The last valid value before the missing value, and the first valid value after the missing value, together provide the basis for the linear interpolation. This is useful if you believe there is local trend in the series.

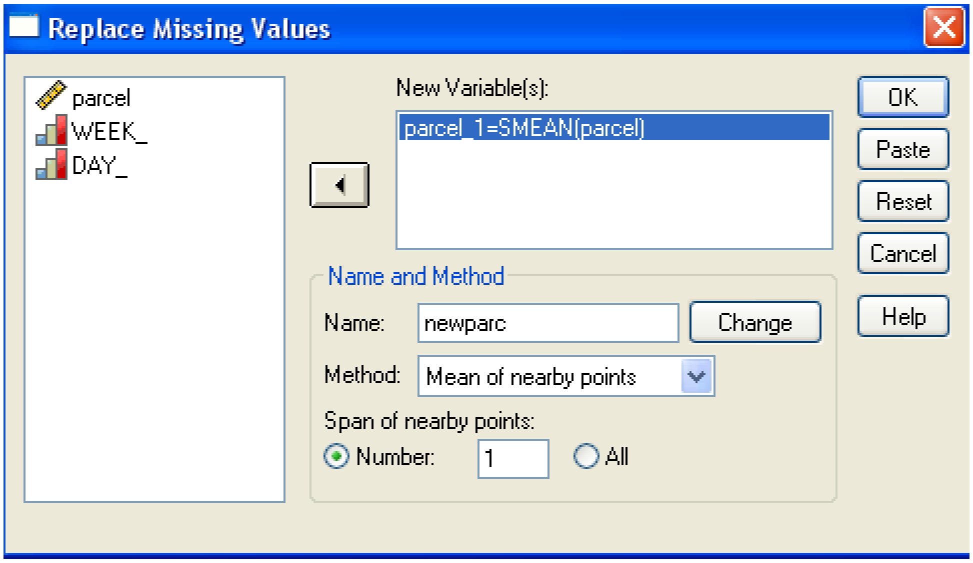
- Finally, the *Linear trend at point* option replaces missing values by running regression on all of the valid data, where the dependent variable is the series itself and the independent variable is time. The regression equation is developed from the observed data and is then used to fit a value for missing observations

  Select the variable **parcel**
  Click the **right arrow** to move it into the New Variables box

By default a new series variable called *parcel_1* will be created with complete values. This name can be changed; we will change the variable name to *newparc*.

Change the Name text to **newparc**
Select **Mean of nearby points** from the Method drop-down list
Change the Span of nearby points: **Number** text box value to **1**
Click **Change**

Figure 2.7 Replace Missing Values Dialog Box



Replace Missing Values

parcel
WEEK_
DAY_

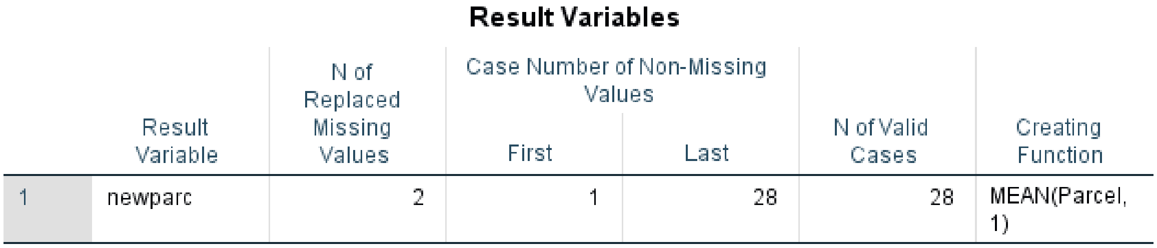
New Variable(s):
parcel_1=SMEAN(parcel)

Name and Method
Name: newparc
Method: Mean of nearby points
Span of nearby points:
● Number: 1
○ All

We changed the default span of nearby points from 2 to 1 because one of the missing values for this series is at case 2. There is only one valid measurement available prior to case 2, of course, that from case 1.

Click on **OK**

Figure 2.8 Information on Missing Values Replaced



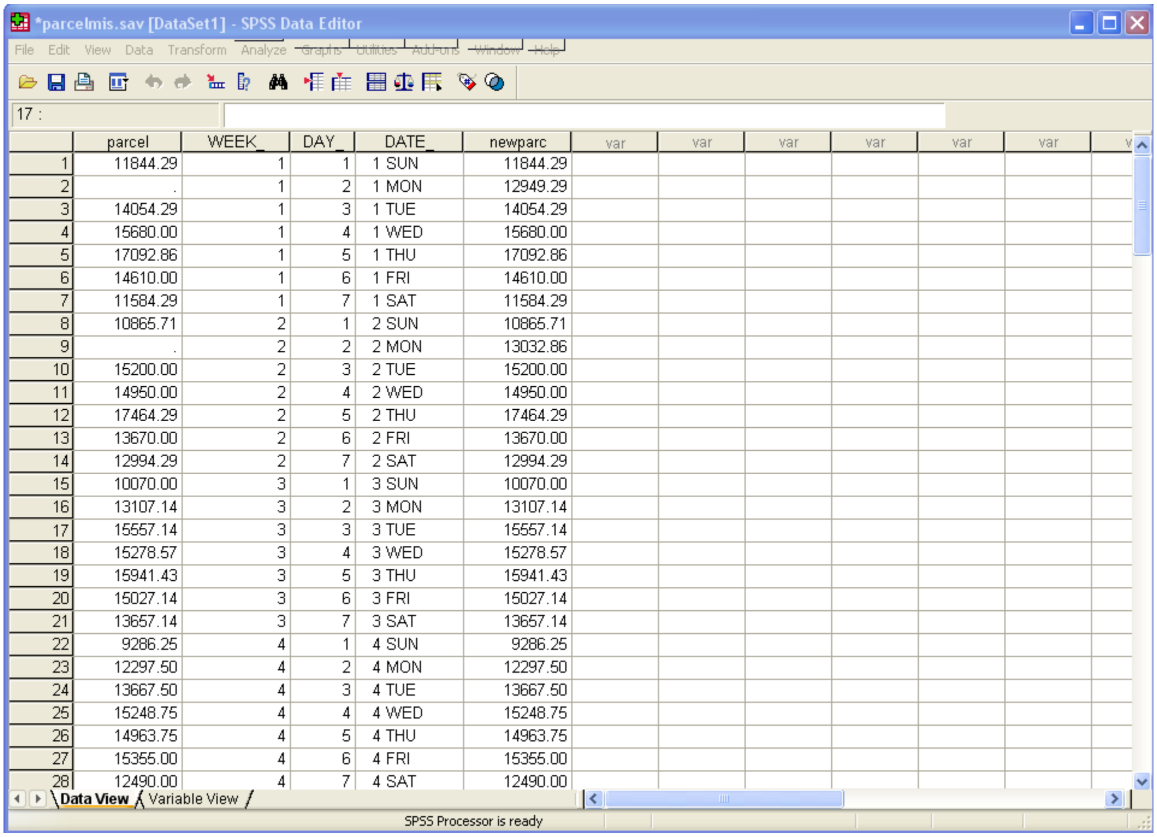
Result Variables

	Result Variable	N of Replaced Missing Values	Case Number of Non-Missing Values		N of Valid Cases	Creating Function
			First	Last		
1	newparc	2	1	28	28	MEAN(Parcel, 1)

In the Output Viewer window a message reports that a new variable called *newparc* has been created, and there is also information on the number of missing values which have been replaced (in this series, two missing values were replaced).

Click the **Data Editor** tool

Figure 2.9 The Newparc Series Displayed in the Data Editor


*parcelmis.sav [DataSet1] - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help
17 :
Data View
Variable View
SPSS Processor is ready

	parcel	WEEK_	DAY_	DATE_	newparc
1	11844.29	1	1	1 SUN	11844.29
2	.	1	2	1 MON	12949.29
3	14054.29	1	3	1 TUE	14054.29
4	15680.00	1	4	1 WED	15680.00
5	17092.86	1	5	1 THU	17092.86
6	14610.00	1	6	1 FRI	14610.00
7	11584.29	1	7	1 SAT	11584.29
8	10865.71	2	1	2 SUN	10865.71
9	.	2	2	2 MON	13032.86
10	15200.00	2	3	2 TUE	15200.00
11	14950.00	2	4	2 WED	14950.00
12	17464.29	2	5	2 THU	17464.29
13	13670.00	2	6	2 FRI	13670.00
14	12994.29	2	7	2 SAT	12994.29
15	10070.00	3	1	3 SUN	10070.00
16	13107.14	3	2	3 MON	13107.14
17	15557.14	3	3	3 TUE	15557.14
18	15278.57	3	4	3 WED	15278.57
19	15941.43	3	5	3 THU	15941.43
20	15027.14	3	6	3 FRI	15027.14
21	13657.14	3	7	3 SAT	13657.14
22	9286.25	4	1	4 SUN	9286.25
23	12297.50	4	2	4 MON	12297.50
24	13667.50	4	3	4 TUE	13667.50
25	15248.75	4	4	4 WED	15248.75
26	14963.75	4	5	4 THU	14963.75
27	15355.00	4	6	4 FRI	15355.00
28	12490.00	4	7	4 SAT	12490.00

Returning to the Data Editor window, the new parcel series is displayed complete with the new values for the first two Mondays of the series.

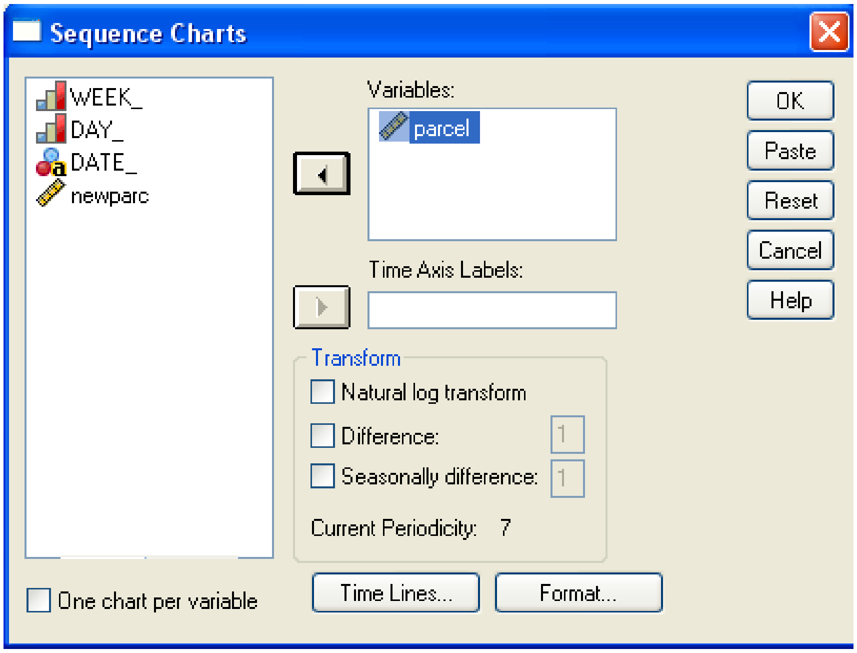
The Sequence Chart

Having collected the relevant data, defined dates, and explored missing data, the next step is to identify any past patterns in the data before deciding upon an appropriate time series technique. Note that you would usually first look at the sequence plot for the original data series with missing data, and then the new series for which missing values have been substituted. To run a sequence chart:

> Click **Analyze -> Forecasting -> Sequence Charts**
> Select the variable **parcel** and move into the Variables box

Note that the current level of periodicity, which was defined through the Define Dates dialog, is shown towards the bottom of the sequence dialog box. Also, if no variable is added to the *Time Axis Label(s)* variable box, the *DATE_* variable will supply labels for the time axis. This is what we desire.

Figure 2.10 Requesting a Sequence Chart for Parcel



Variables: parcel

Time Axis Labels:

□ Natural log transform

□ Difference: 1

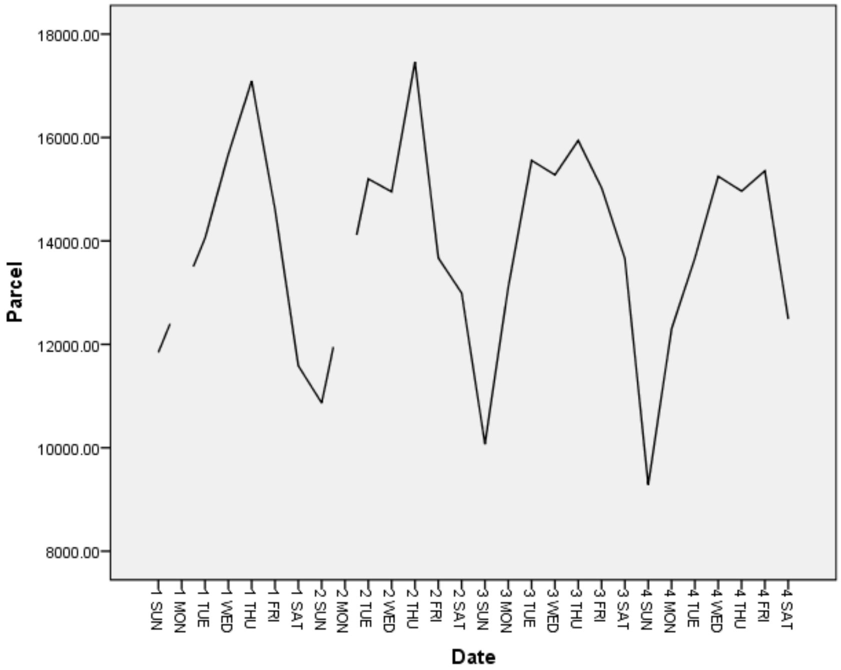
□ Seasonally difference: 1

Current Periodicity: 7

□ One chart per variable

Click **OK**

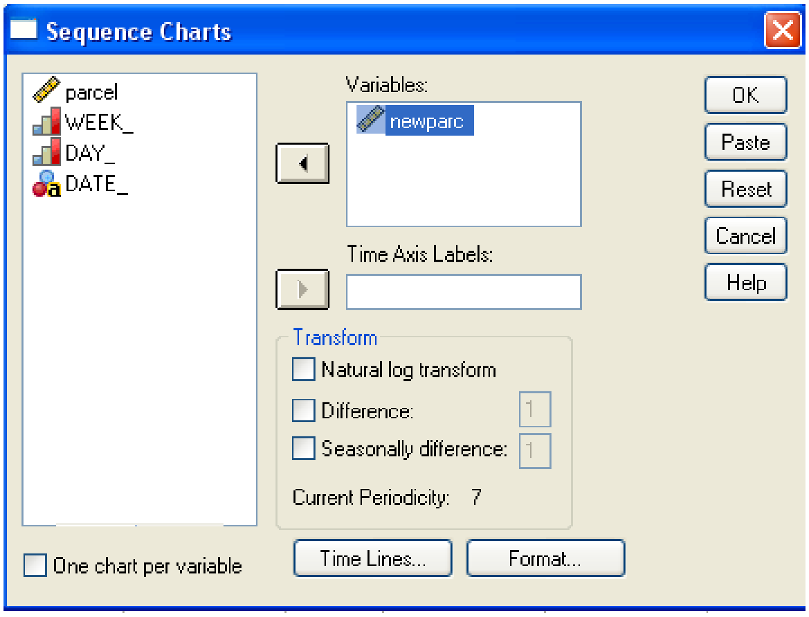
Figure 2.11 Sequence Chart for Parcel



Early in the series, we can see gaps resulting from the two missing values for *parcel*. From the sequence plot, you can often make a preliminary assessment as to whether trend and seasonal components are needed to fit the series.

Click on the **Dialog Recall tool** , and then click **Sequence Charts**
Highlight the variable **parcel** in the Variables box
Click the **arrow** to move parcel back to the left box
Select the variable **newparc** and move it into the Variables box

Figure 2.12 Requesting a Sequence Chart for newparc



Variables: newparc

Time Axis Labels:

Transform

□ Natural log transform

□ Difference: 1

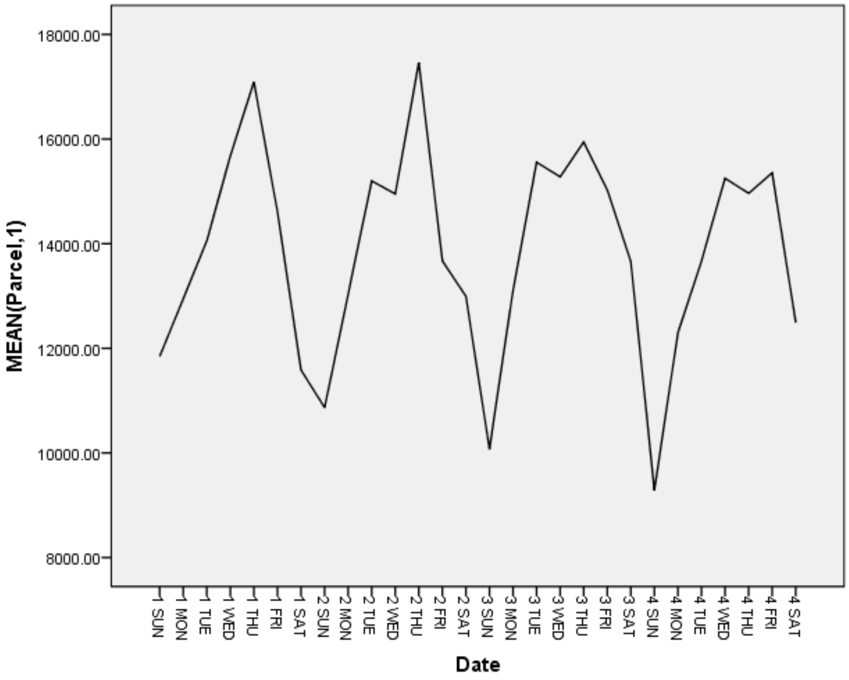
□ Seasonally difference: 1

Current Periodicity: 7

□ One chart per variable

Click **OK**

Figure 2.13 Sequence Chart for newparc



Notice that in the second sequence chart there are no gaps since we have replaced the missing data with the mean of nearby points.

Summary

In this chapter we have:

- Defined periodicity and shown the Define Dates procedure in SPSS
- Discussed some data requirements for time series analysis
- Shown how to deal with missing values