Handout For:
Automated Real-Time Forecasting of Stream Conditions with SAS®
Detailed Instructions and Documentation

Samuel T. Croker, Independent Consultant
Shane L. Hornibrook, Independent Consultant
Tomonori Ishikawa, USC Department of Statistics, Columbia, SC

Contents

OBTAINING AND PREPARING THE STREAM DATA

Short period data is available from the Water Resources section of the United States Geological Survey (USGS) online. The USGS has graciously allowed queryable access of realtime data collected from thousands of automatic monitoring stations from across the country. A document containing some basic information about http queries is located at
`http://waterdata.usgs.gov/nwis/news/?automated_retrieval_info`.

The in the PDF version of this document, all URLs are clickable to view the results of the web queries.

Parameter Codes: `http://nwis.waterdata.usgs.gov/usa/nwis/pmcodes`

SITE INVENTORY

It is possible to query the site inventory to get site names, geographic coordinates and many other variables using the /inventory.

`http://waterdata.usgs.gov/nwis/inventory?multiple_site_no=01581920,01582000,01595000,01609000,01596500,01597500,01598500,01649150,03075500,03076500,01591000,03078000,03079000&format=rdb&column_name=agency_cd&column_name=site_no&column_name=station_nm`

DATA EXTRACTION

The following query is one of the ways to get the observation data. For multiple site numbers, this file will have header information that has to be filtered out. Also, the columns are not guaranteed to be in the same place so much care must be taken. For the sake of simplicity, only one parameter is loaded in this example. This is done by using the `index_pmcode_00065=1` statement, which corresponds to the `STAGE` parameter.

For these multiple sites, only the STAGE parameter is common to all.

Multiple Site Data: http://waterdata.usgs.gov/nwis/uv?period=31&multiple_site_no=01581920,01582000,01595000,01609000,01596500,01597500,01598500,01649150,03075500,03076500,01591000,03078000,03079000&format=rdb&index_pmcode_00065=1

URL Access Method of Filename Statement The URL access method associates the target of a url with a SAS FILEREF. The ampersands that are embedded in the URL are separators for the different parameters that are passed to the NWIS web server and must be masked using the %str() masking function. This function will prevent SAS from tokenizing the ampersands as a macro variable start point and send them along as actual ampersands.

```
filename maryland url
    "http://waterdata.usgs.gov/nwis/uv?
     period=31
     %str(&)multiple_site_no=01581920,01582000,...,03079000
     %str(&)format=rdb";
     options datestyle=ydm;
```

Preparing the Data The text file that lies behind the url for the data has a lot of header information that has to be filtered out. This header information could be used to automate the extraction process but in this example it is assumed that the data layout is known. An excerpt from the file is shown below. This represents the header information that is inserted when the stream switches from one site to another.

```
...
USGS 01581920 2007-11-07 06:00 1.43
USGS 01581920 2007-11-07 06:15 1.43
#
# Data provided for site 01582000
#    DD parameter   Description
#    02   00065     Gage height, feet
#
agency_cd site_no datetime 02_00065 02_00065_cd
5s 15s 16d 14n 10s
USGS 01582000 2007-10-07 00:00 0.17
USGS 01582000 2007-10-07 00:14 0.17
...
```

These files are pretty well set up for automated extraction. For all sites in this extract the data rows are identified by the first four columns having an agency_cd of USGS. The data is tab delimited, defined by the URL parameter format=rdb. The date and time variable are only separated by a space so they are parsed using the SUBSTR function.

```
options datestyle=ydm;
data western_maryland_data;
    infile maryland end=eof dlm='09'x dsd; /* TAB DELIMITED DATA */
    format datestamp datetime16.;
    length agency_cd $5 site_no $30 flow 8 stage 8 precip 8;
    input agency_cd @;
    if agency_cd~='USGS' then delete;
    else input site_no $ dtm $16. fil1 $ stage ;
    date=input(substr(dtm,1,10),anydtdte10.);
    time=input(substr(dtm,12,5),anydttme5.);
    datestamp=dhms(date,hour(time),minute(time),0);
    drop fil1 date time;
run;
```

Correcting Anomalies There are often a number of anomalies in the data that comes from NWIS. These are not necessarily errors but do present a problem for time series analysis and forecasting techniques.

```
agency_cd site_no datetime 02_00065 02_00065_cd
5s 15s 16d 14n 10s
USGS 01582000 2007-10-07 00:00 0.17
USGS 01582000 2007-10-07 00:14 0.17
USGS 01582000 2007-10-07 00:15 0.17
USGS 01582000 2007-10-07 00:30 0.17
```

In this case, the quarterly hour observations also have a 14 minute shadow. This may be a valid observation but causes a difficulty due to the fact that the observations are not equally spaced. Another problem can occur when some sites in the comparison list have quarterly hour data and others have only hourly data. To solve both of these issues, and to account for missing values, `PROC TIMESERIES` is employed.

```
proc timeseries data=western_maryland_data out=working;
by site_no;
id datestamp interval=hour accumulate=max;
var stage;
run;
```

The `ID` statement is the crucial statement in this transformation. Above we are accumulating the observations to the hourly level and taking the maximum of the other observations. The maximum makes sense for this data but for other data it might be better to use the sum or average. This step can also be done directly in the HPFENGINE procedure but it is also nice to have prepared data set aside for other uses so doing this during pre-processing makes sense.

FORECASTING WITH HPF

Forecasting using the SAS High Performance Forecasting System is a bit of overkill for these series, but it is a reasonable example of how it is done.

```
%macro buildarimaspecs;
    %let mdl=0;
    %do p=0 %to 3;
        %do q=0 %to 3;
            %do d=0 %to 2;
                %if &d=0 %then %let dif=0;
                %if &d=1 %then %let dif=1;
                %if &d=2 %then %let dif=7;

                %let mdl=%eval(&mdl+1);
                proc hpfarimaspec repository=work.arima name=amd&mdl;
                    forecast symbol=stage transform=none p=&p dif=&dif q=&q;
                    estimate method=ml;
                run;
                proc hpfarimaspec repository=work.arima name=bmd&mdl;
                    forecast symbol=stage transform=boxcox(0.5) p=&p dif=&dif q=&q;
                    estimate method=ml;
                run;
                proc hpfarimaspec repository=work.arima name=cmd&mdl;
                    forecast symbol=stage transform=none noint p=&p dif=&dif q=&q;
                    estimate method=ml;
                run;
```

```
                proc hpfarimaspec repository=work.arima name=dmd&mdl;
                    forecast symbol=stage transform=boxcox(0.5) noint p=&p dif=&dif q=&q;
                    estimate method=ml;
                run;
            %end;
        %end;
    %end;

%mend buildarimaspecs;

proc catalog catalog=work.arima kill; run;quit;

%buildarimaspecs;

proc catalog catalog=work.arima; contents out=speccont; run; quit;

proc sql noprint;
    select distinct name into :spec separated by ' ' from speccont;
quit;

%put &spec;

proc hpfselect repository=work.arima
         name=myselect
         label="My Selection List";
    select criterion=mape holdout=72;
    spec &spec ;
run;

%let interval=hour;
%let back=12;
%let lead=24;
ods trace on;
proc hpfdiag data=maryland
        print=all
        repository=work.arima
        criterion=mape
        back=&back
        lead=&lead
        outest=diagest;
    by site_no;
    id datestamp interval=hour accumulate=max;
    forecast stage;
    esm;
    arimax outlier=(detect=maybe) method=minic;
    trend dif=auto;
    transform type=auto;
run;
ods trace off;

ods output modelselection=mdlselect parameterestimates=estimates;

proc hpfengine
        repository=work.arima
        inest=diagest
        data=maryland
```

```
        outfor=outfor
        outest=outest
        back=&back
        lead=&lead
        print=(select estimates);
    by site_no;
    id datestamp interval=hour accumulate=avg;
    forecast stage;
 run;
proc sort data=mdlselect; by statistic;run;
```

REFERENCES

Box, George E.P., Gwilym M. Jenkins and Gregory C. Reinsel. 1994. *Time Series Analysis: Forecasting and Control*, 3rd ed. Upper Saddle River, NJ: Prentice-Hall.

Brocklebank, John and David A. Dickey. 2003. *SAS® for Forecasting Time Series*, 2nd ed. Cary, NC: SAS Institute Inc.

Gelso, Charlie, Larry Coburn. 2006. *Guide to Maryland Trout Fishing: The Catch and Release Streams* Carter, OK: Falling Star Publishing

Cartier, Jeff. "The Power of the Graphics Template Language." *Proceedings of the 30th Annual SAS® Users Group International Conference*. April 2004.
<http://support.sas.com/rnd/datavisualization/papers/sugi30/GTL.pdf>
(Accessed July 18, 2007).

Croker, Samuel T. "Effective Forecast Visualization with SAS/GRAPH." *SAS Global Forum 2007 Proceedings*. April 2007.
<http:
//www8.sas.com/scholars/Proceedings/2006/DataPresentation/DP01_06.PDF>

Shumway, Robert H. and David S. Stoffer. 2006. *Time Series Analysis and Its Applications with R Examples*, 2nd ed. New York: Springer Science+Business Media, LLC.

CONTACT INFORMATION

We value and encourage your comments and questions! You can find the latest version of the SAS code for this paper at: `http://www.scoyote.net/forecasting/`. Please note that we may update this code for use in other papers.

You can contact the authors at:

Name: Samuel T. Croker
E-Mail: `scoyote at scoyote.net`
Web: `http://www.scoyote.net/forecasting/`

Name: Shane L. Hornibrook
E-Mail: `sesug_paper at shanehornibrook.com`

Name: Tomonori Ishikawa
E-Mail: `ish at alum.mit.edu`
Web: `http://www.stat.sc.edu/~ishikawa/`