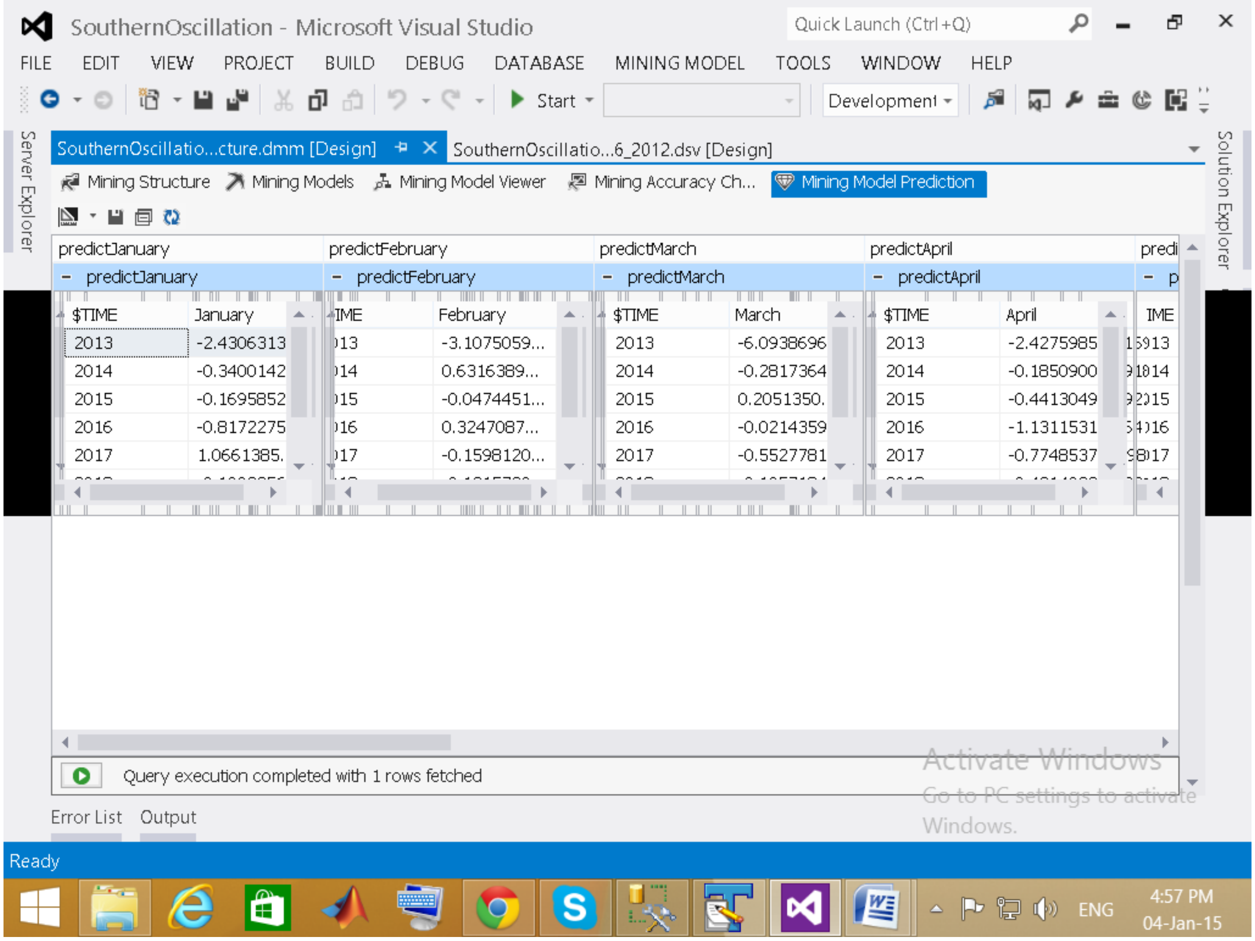
<u>Flashbacking on previous task report:</u>

In the last task report submitted on 04.01.2015 I reported that, I downloaded data set from http://www.esrl.noaa.gov/psd/gcos_wgsp/Timeseries/SOI/ for analyzing time series and forecasting. This data is about normalized surface pressure difference between Tahiti and Darwin islands for every month of the year starting from 1866 to 2012.

Then I inserted the data set into database in a table named “soi1866_2012” which contains columns named "year", "january", "february", "march", "april", "may", "june", "july", "august", "september", "october", "november", " december".

After that I made forecast for next 20 steps. Forecasts were made for every month of the year separately. For example, for April, I made forecast on the point that what amount of surface oscillation can occur in April 2013, April 2014 … … April 2032. I used one data for each April from the year 1886 to 2012 as training data set. In this way I made forecast for other months also.

Here forecasts are for next 20 steps.



predictJanuary

$TIME	January
2013	-2.4306313
2014	-0.3400142
2015	-0.1695852
2016	-0.8172275
2017	1.0661385.

predictFebruary

IME	February
13	-3.1075059...
14	0.6316389...
15	-0.0474451...
16	0.3247087...
17	-0.1598120...

predictMarch

$TIME	March
2013	-6.0938696
2014	-0.2817364
2015	0.2051350.
2016	-0.0214359
2017	-0.5527781

predictApril

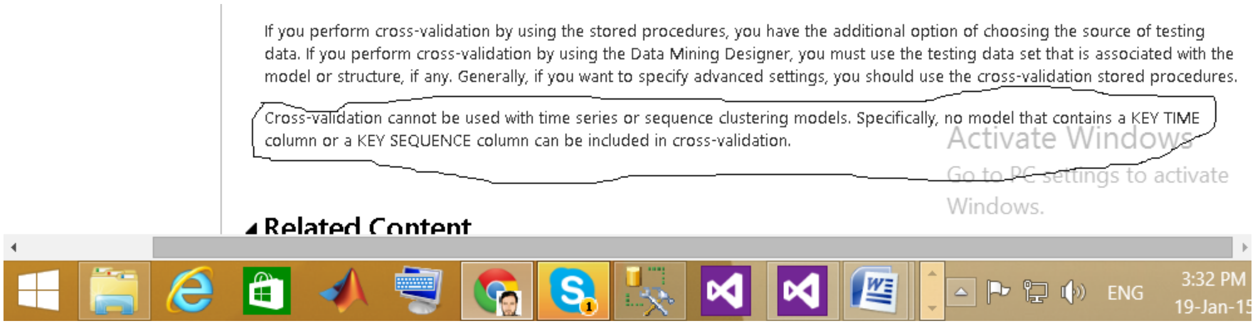
$TIME	April
2013	-2.4275985
2014	-0.1850900
2015	-0.4413049
2016	-1.1311531
2017	-0.7748537

Query execution completed with 1 rows fetched

I said you that I was planning to read documentation http://msdn.microsoft.com/en-us/library/bb895174.aspx on ‘Cross Validation’ to meaesure the accuracy of mining models.

New report

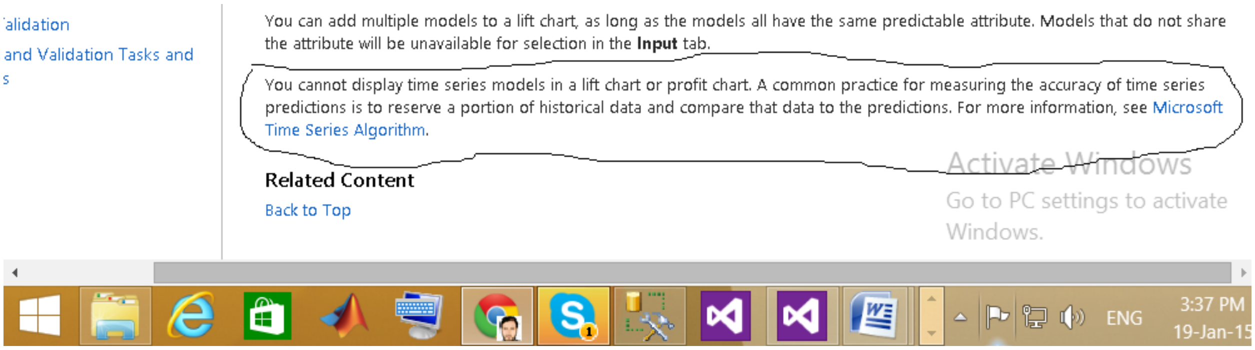
I read whole documentation on “Cross validation” and partly on “Lift Chart” , “Classification Matrix”. But none of these data model accuracy measuring tools is not valid for measuring accuracy of time series models.



If you perform cross-validation by using the stored procedures, you have the additional option of choosing the source of testing data. If you perform cross-validation by using the Data Mining Designer, you must use the testing data set that is associated with the model or structure, if any. Generally, if you want to specify advanced settings, you should use the cross-validation stored procedures.

Cross-validation cannot be used with time series or sequence clustering models. Specifically, no model that contains a KEY TIME column or a KEY SEQUENCE column can be included in cross-validation.

Related Content



alidation
and Validation Tasks and
s

You can add multiple models to a lift chart, as long as the models all have the same predictable attribute. Models that do not share the attribute will be unavailable for selection in the **Input** tab.

You cannot display time series models in a lift chart or profit chart. A common practice for measuring the accuracy of time series predictions is to reserve a portion of historical data and compare that data to the predictions. For more information, see Microsoft Time Series Algorithm.

Related Content

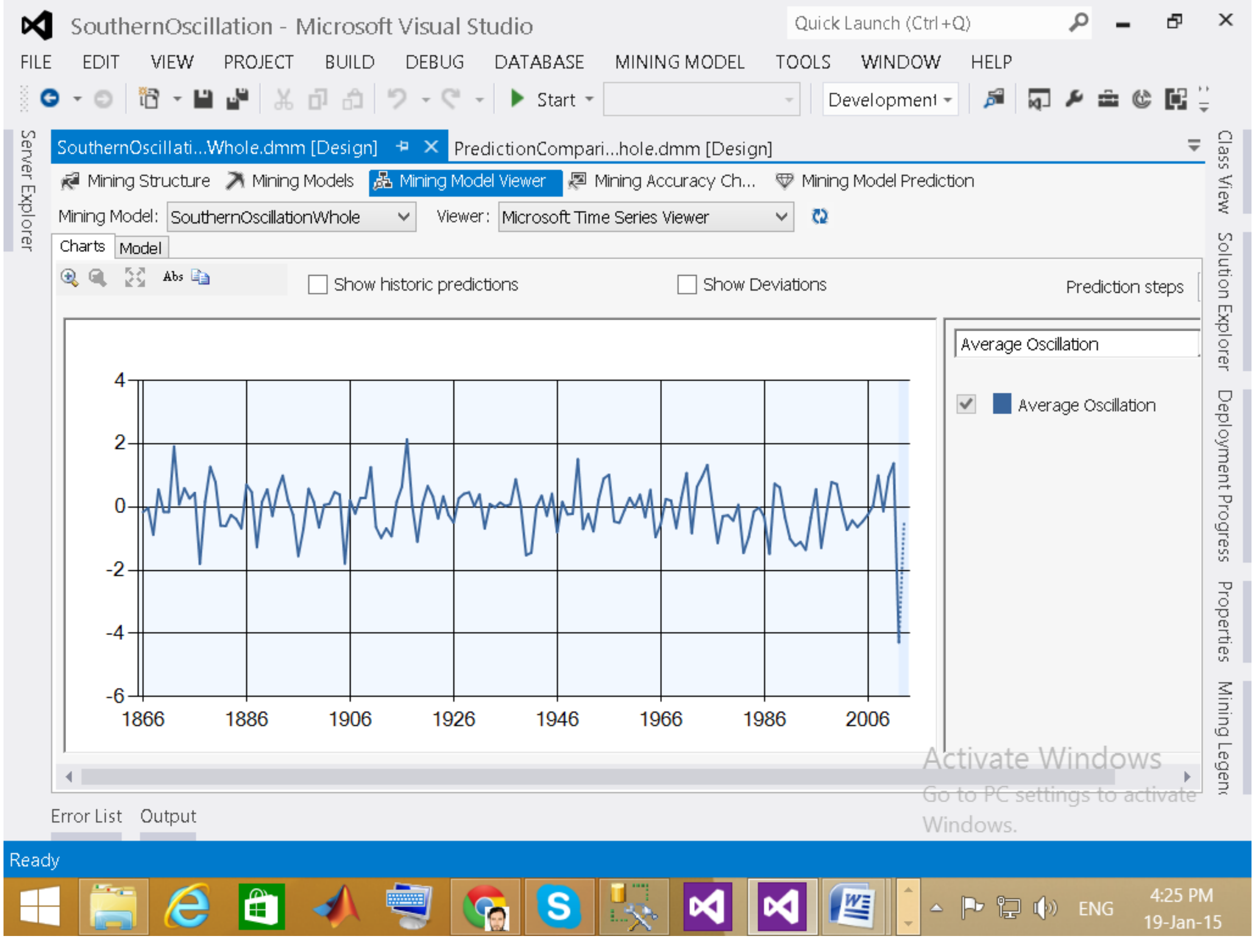
Back to Top

After that I decided to compare two predictions done by the same model with different amount of traning data to measure accuracy of the model according to this documentation http://msdn.microsoft.com/en-us/library/cc879295.aspx . For doing this I needed to make a cross prediction according to this documentation http://msdn.microsoft.com/en-us/library/cc879284.aspx and after that I compared original prediction which I have done and showed you in the previous task report with cross prediction results.

Step 01: Making a cross prediction

The process of using data from one series to predict trends in another series is called cross-prediction. That means a model will be trained from a series but will predict for another series. For example, here I trained a time series model named "SouthernOscillationWhole" by the data in the series named "SouthernWholeOscillation" which includes (whole data set) surface pressure occurred in all months (Jan - Dec) from the year 1866 to 2012.

This is model view of the model "SouthernOscillationWhole" trained by the data in series "SouthernWholeOscillation". **The prediction part here shown by dotted line is not important to consider**. This graph is only to observe the data fluctuation throughout the year from 1866 to 2012.

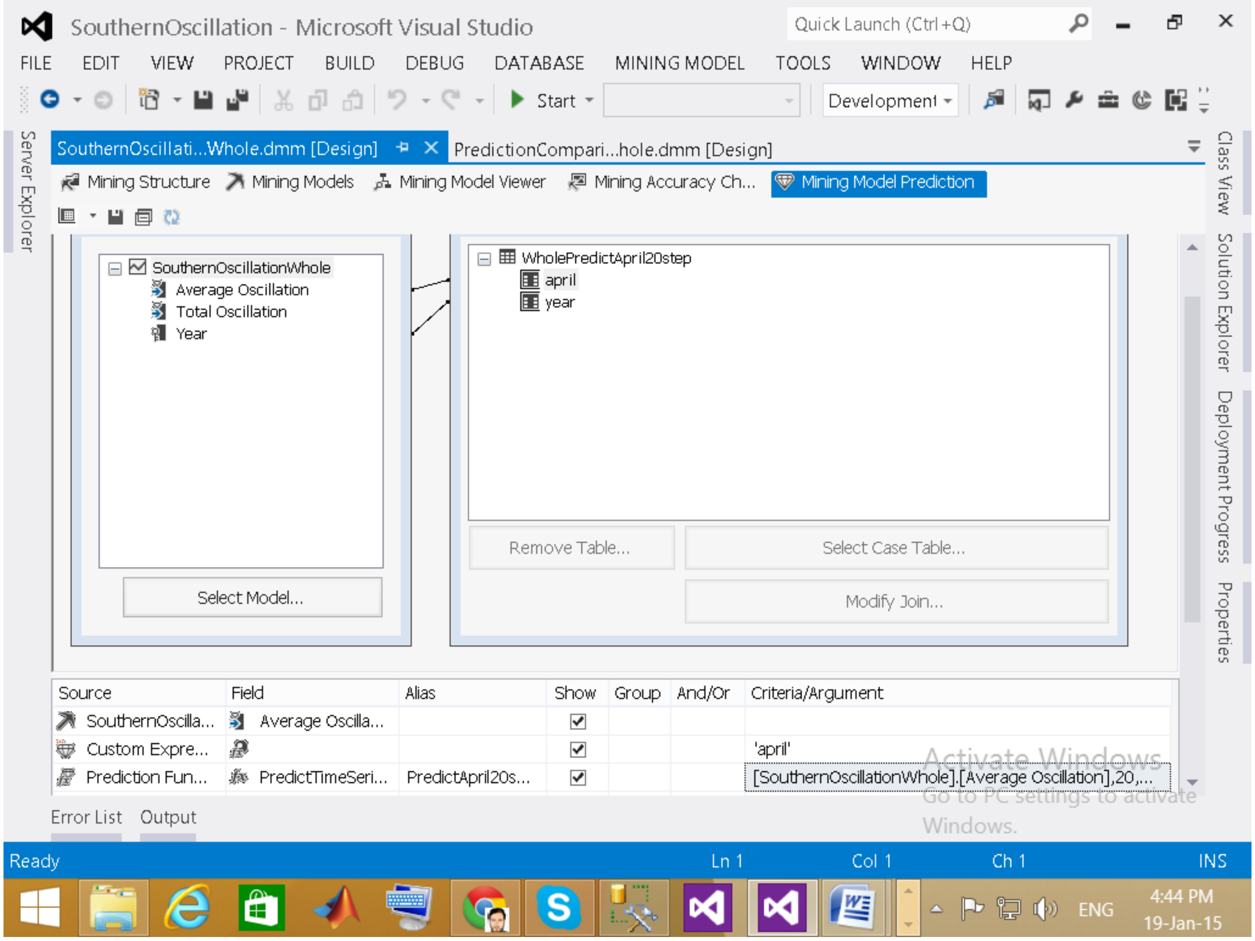

Mining Model: SouthernOscillationWhole

Viewer: Microsoft Time Series Viewer

☐ Show historic predictions

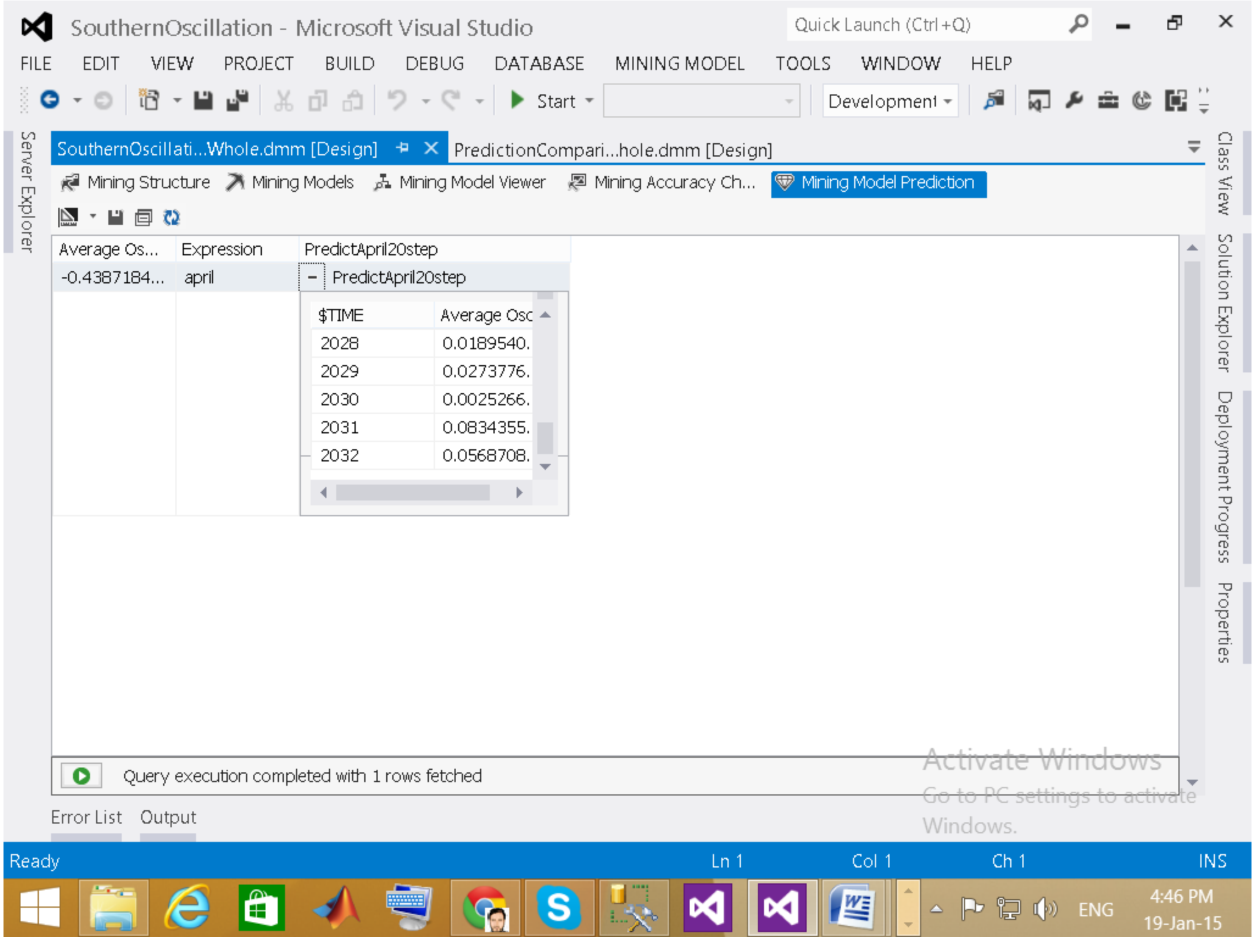
☐ Show Deviations

After training from these data I used this trained model to predict for the series named “SeriesApril” which includes one data for each April from the year 1886 to 2012. This is called cross prediction for month of April.



Source	Field	Alias	Show	Group	And/Or	Criteria/Argument
SouthernOscilla...	Average Oscilla...		☑			
Custom Expre...			☑			'april'
Prediction Fun...	PredictTimeSeri...	PredictApril20s...	☑			[SouthernOscillationWhole].[Average Oscillation],20,...

Here in the following picture the result of cross prediction for from the month of April of 2013 to April of 2032. In this way I made cross prediction for other months also.



Average Os...	Expression	PredictApril20step
-0.4387184...	april	− PredictApril20step

$TIME	Average Osc
2028	0.0189540.
2029	0.0273776.
2030	0.0025266.
2031	0.0834355.
2032	0.0568708.

Query execution completed with 1 rows fetched

Then I saved the result of cross prediction for April in a table named “WholePredictApril20step” in the database named “SouthernOscillation” which I attached with the mail for your kind consideration.

Step 02: Comparing the cross prediction result and original predicrtion result:

Now summary is following:

In the database “SouthernOscillation” contains two prediction result table for the month of April named “PrimaryPredictApril20step” and “WholePredictApril20step”

The table “PrimaryPredictApril20step” contains prediction result for the month of April done the model named “SouthernOscillationModel” trained by the data in the series named “SeriesApril” which includes one data for each April from the year 1886 to 2012 to predict surface pressure in April. I have done it and showed you in the previous task report. The table “PrimaryPredictApril20step” contains two columns named “PredictApril.$TIME”, “PredictApril.April”.

The table “WholePredictApril20step” contains cross prediction result for the month of April done by the model “SouthernOscillationWhole” which was trained by the data in the series named “SouthernWholeOscillation” which includes (whole data set) surface pressure occurred in all months (Jan - Dec) from the year 1866 to 2012. The table “WholePredictApril20step” contains two columns named “PredictAprilWhole.$TIME”, “PredictAprilWhole.Average Oscillation”.

It is important to mention that, for the model “SouthernOscillationWhole” I used the same algorithm (Time series) and algorithm parameters which I used for the model named “SouthernOscillationModel”. So that the prediction results tables “WholePredictApril20step” and “PrimaryPredictApril20step” of two models respectively can be considered the prediction of same model but by different training data. But for the easiness of task I used the model using different names.

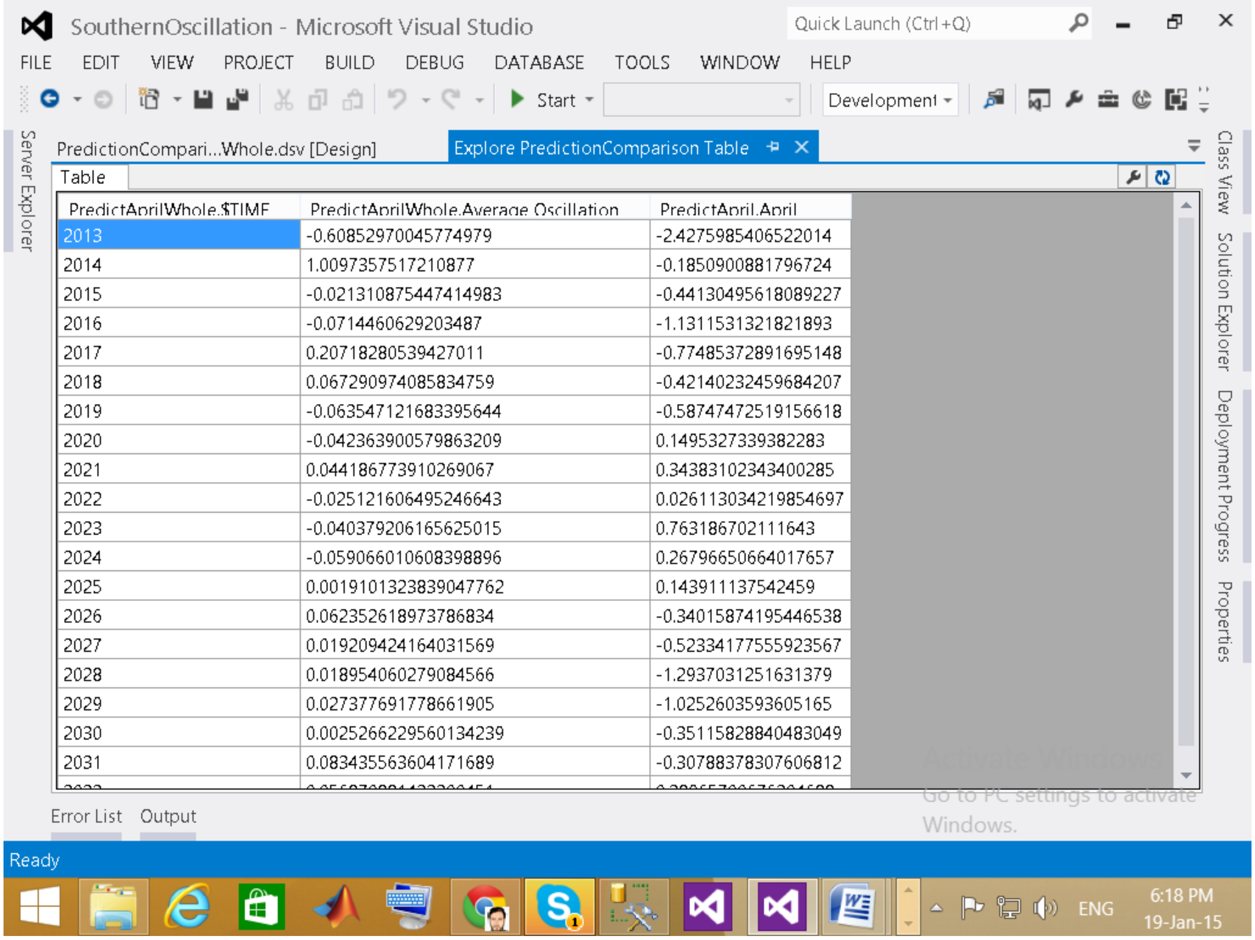
Comparison:

I made another data source view named “PredictionComparisonPrimaryVsWhole” in the project SouthernOscillation in MS data analysis tools using the following sql

```
select
[dbo].[WholePredictApril20step].[PredictAprilWhole.$TIME],[dbo].[WholePredictApril20step]
.[PredictAprilWhole.Average
Oscillation],[dbo].[PrimaryPredictApril20step].[PredictApril.April] from
[dbo].[WholePredictApril20step],[dbo].[PrimaryPredictApril20step] where
[dbo].[WholePredictApril20step].[PredictAprilWhole.$TIME]=[dbo].[PrimaryPredictApril20ste
p].[PredictApril.$TIME]
```

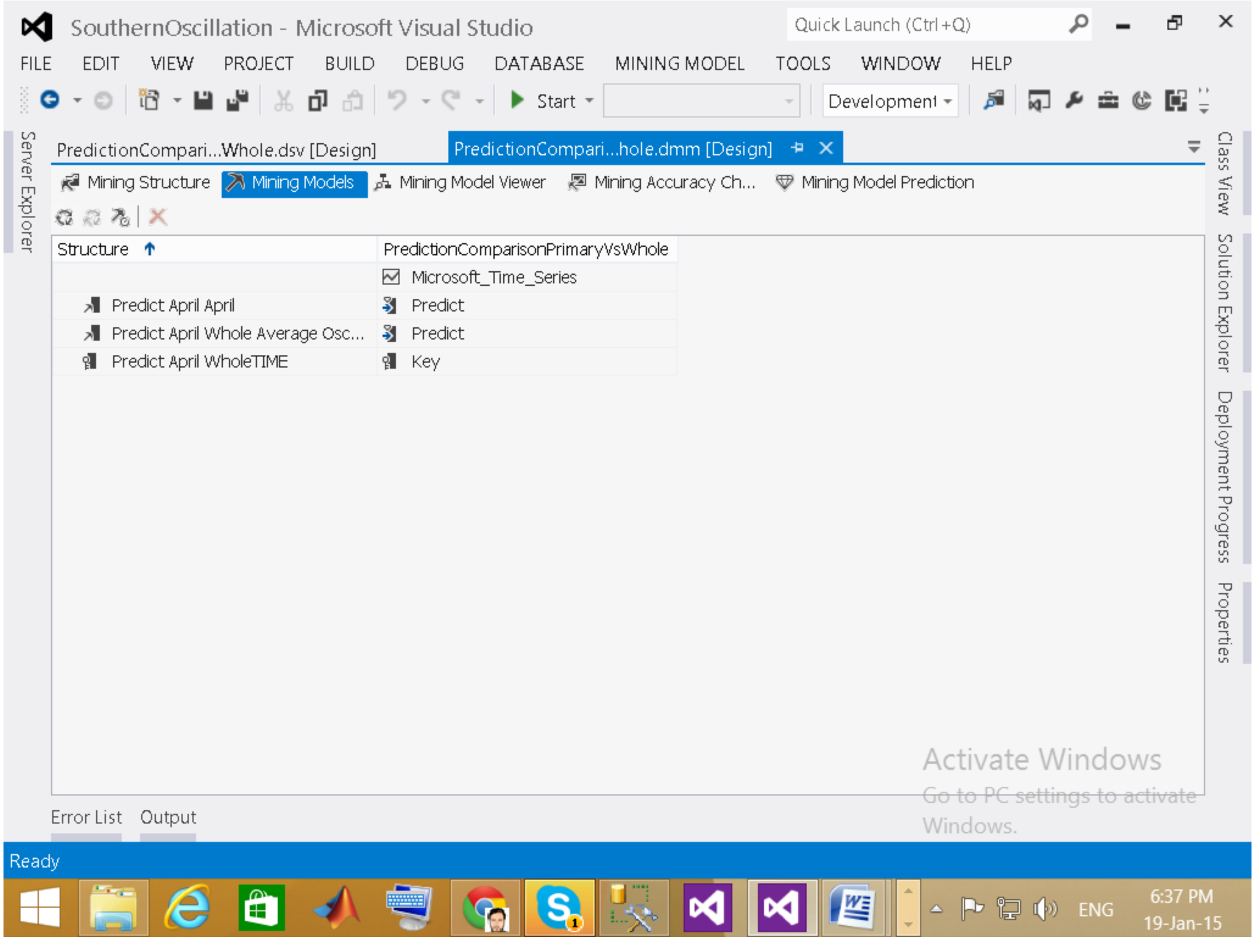
That means new data source view contains three columns named “`PredictAprilWhole.$TIME`”, “`PredictAprilWhole.Average Oscillation`”, “`PrimaryPredictApril20step`” where the columns “`PredictAprilWhole.Average Oscillation`” and “`PrimaryPredictApril20step`” contain the cross prediction result (done by the model “SouthernOscillationWhole”) and primary prediction result (done by the model “SouthernOscillationModel”) respectively.

The following picture is the explored view of the data source view named “PredictionComparisonPrimaryVsWhole”


SouthernOscillation - Microsoft Visual Studio
Quick Launch (Ctrl+Q)
FILE EDIT VIEW PROJECT BUILD DEBUG DATABASE TOOLS WINDOW HELP
Start
Development
Server Explorer
PredictionCompari...Whole.dsv [Design]
Explore PredictionComparison Table
Table
Class View
Solution Explorer
Deployment Progress
Properties
Activate Windows
Go to PC settings to activate Windows.
Error List
Output
Ready
ENG
6:18 PM
19-Jan-15

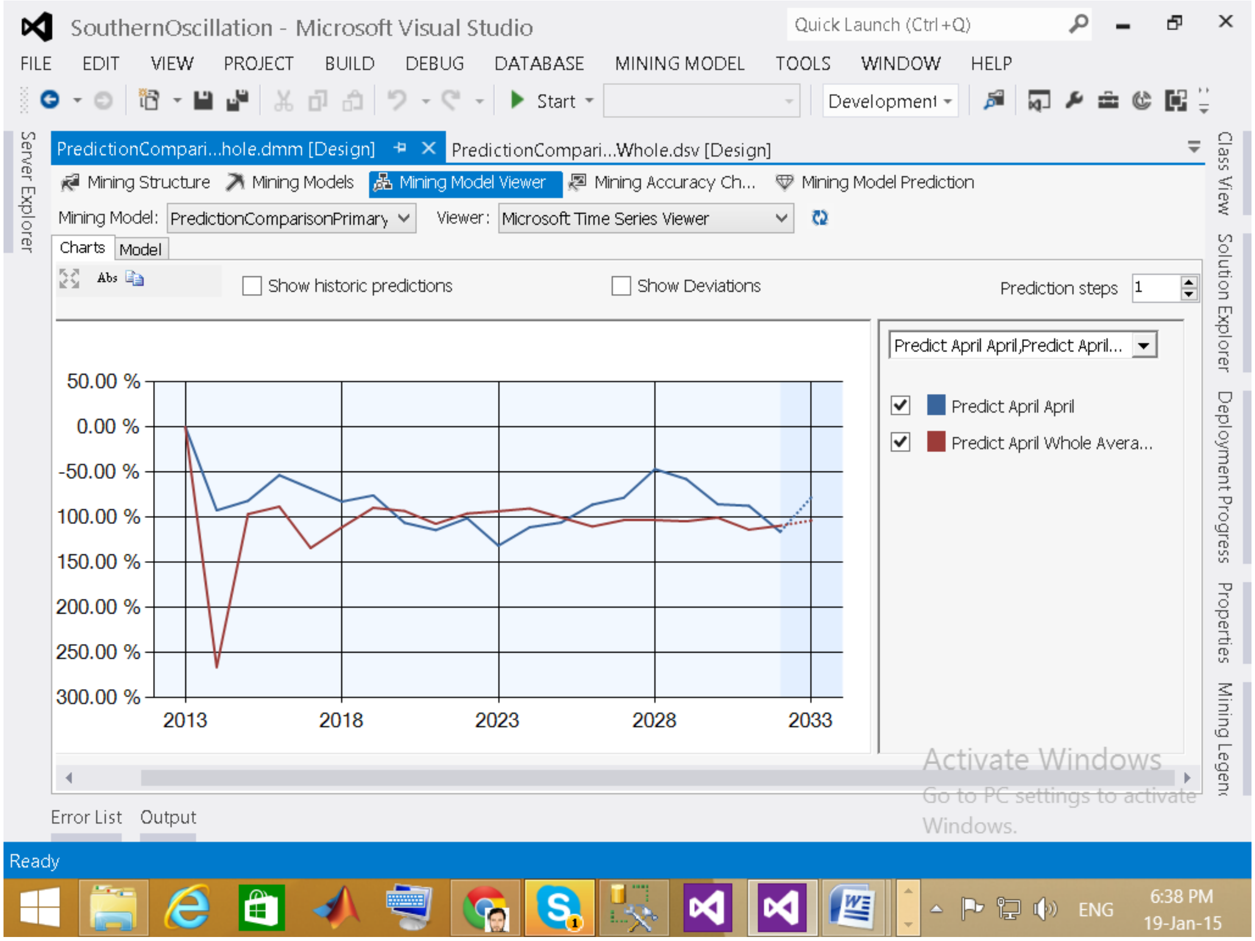
PredictAprilWhole.$TIME	PredictAprilWhole.Average Oscillation	PredictApril.April
2013	-0.60852970045774979	-2.4275985406522014
2014	1.0097357517210877	-0.1850900881796724
2015	-0.021310875447414983	-0.44130495618089227
2016	-0.0714460629203487	-1.1311531321821893
2017	0.20718280539427011	-0.77485372891695148
2018	0.067290974085834759	-0.42140232459684207
2019	-0.063547121683395644	-0.58747472519156618
2020	-0.042363900579863209	0.1495327339382283
2021	0.044186773910269067	0.34383102343400285
2022	-0.025121606495246643	0.026113034219854697
2023	-0.040379206165625015	0.763186702111643
2024	-0.059066010608398896	0.26796650664017657
2025	0.0019101323839047762	0.143911137542459
2026	0.062352618973786834	-0.34015874195446538
2027	0.019209424164031569	-0.52334177555923567
2028	0.018954060279084566	-1.2937031251631379
2029	0.027377691778661905	-1.0252603593605165
2030	0.0025266229560134239	-0.35115828840483049
2031	0.083435563604171689	-0.30788378307606812

Then I created another time series model named “PredictionComparisonPrimaryVsWhole” using the the column “`PredictAprilWhole.$TIME`” as key time and the columns “`PredictAprilWhole.Average Oscillation`” and “`PrimaryPredictApril20step`” as input and predict type. It is impotant to mention the purpose of creating of this model named “PredictionComparisonPrimaryVsWhole” is not to predict but to observe the graph of model view (which is actualy a visualized graph of our cross prediction and primary prediction) so that our conception about the difference between cross prediction and primary prediction.



Structure	PredictionComparisonPrimaryVsWhole
	Microsoft_Time_Series
Predict April April	Predict
Predict April Whole Average Osc...	Predict
Predict April WholeTIME	Key

This is the model view of the model named “PredictionComparisonPrimaryVsWhole”. **The dotted line indicates prediction which is not impotant to cosider**. The plain line with blue colour shows primary prediction for the month of April from the year 2013 to 2032 and the plain line with red colour shows the cross prediction for the month of April from the year 2013 to 2032. It is the comparison. From this graph we can understand how the prediction done by the same model differes for the different training data.



Mining Model: PredictionComparisonPrimary

Viewer: Microsoft Time Series Viewer

□ Show historic predictions

□ Show Deviations

Prediction steps 1

Predict April April,Predict April...

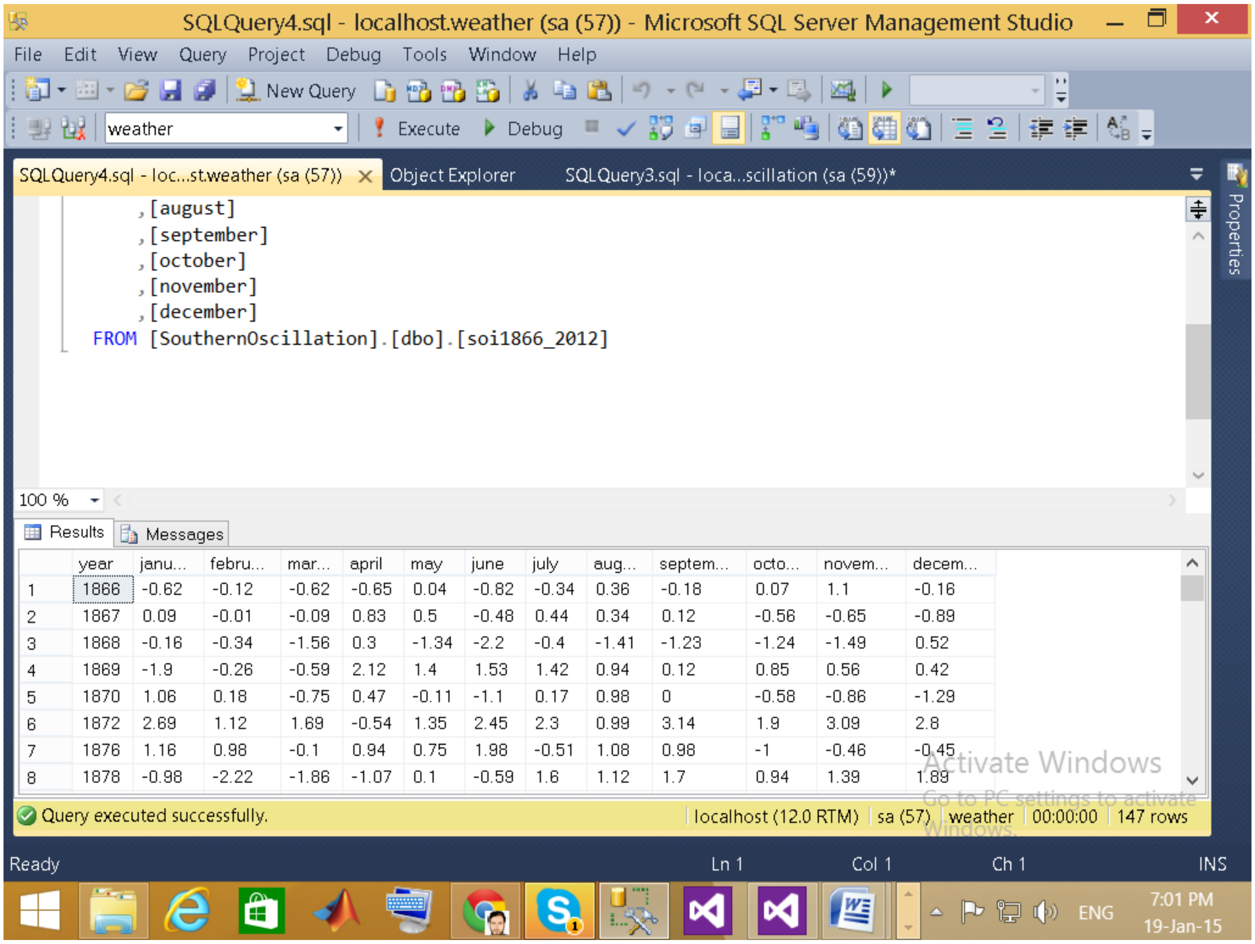
☑ 🟦 Predict April April

☑ 🟥 Predict April Whole Avera...

Ready

Conclusion:

We see in the previous picture that, in 2014, there is enough difference between two prediction while the difference is smaller in follwing years. Why so much large difference in 2014? If we observe the original data we see that it is normal that sometimes surface pressure differes largely between the neighbouring times. For example, in the April 1869 surface pressure is 2.12 while at the same time in 1870 it is only 0.47 . Again in June 1884 surface pressure is 0.98 while it is in july in the same year is -0.28. **When we use the whole data set as training data set for cross prediction, the model faces more fluctuation than when use a part of data as training data set for primary prediction. That is the cause, why cross prediction shows more deeper fluctuation than than the primary prediction.**



SQLQuery4.sql - localhost.weather (sa (57)) - Microsoft SQL Server Management Studio
File Edit View Query Project Debug Tools Window Help
New Query
weather
Execute
Debug
SQLQuery4.sql - loc...st.weather (sa (57))
Object Explorer
SQLQuery3.sql - loca...scillation (sa (59))*
Properties
100 %
Results
Messages

```
      ,[august]
      ,[september]
      ,[october]
      ,[november]
      ,[december]
  FROM [SouthernOscillation].[dbo].[soi1866_2012]
```

	year	janu...	febru...	mar...	april	may	june	july	aug...	septem...	octo...	novem...	decem...
1	1866	-0.62	-0.12	-0.62	-0.65	0.04	-0.82	-0.34	0.36	-0.18	0.07	1.1	-0.16
2	1867	0.09	-0.01	-0.09	0.83	0.5	-0.48	0.44	0.34	0.12	-0.56	-0.65	-0.89
3	1868	-0.16	-0.34	-1.56	0.3	-1.34	-2.2	-0.4	-1.41	-1.23	-1.24	-1.49	0.52
4	1869	-1.9	-0.26	-0.59	2.12	1.4	1.53	1.42	0.94	0.12	0.85	0.56	0.42
5	1870	1.06	0.18	-0.75	0.47	-0.11	-1.1	0.17	0.98	0	-0.58	-0.86	-1.29
6	1872	2.69	1.12	1.69	-0.54	1.35	2.45	2.3	0.99	3.14	1.9	3.09	2.8
7	1876	1.16	0.98	-0.1	0.94	0.75	1.98	-0.51	1.08	0.98	-1	-0.46	-0.45
8	1878	-0.98	-2.22	-1.86	-1.07	0.1	-0.59	1.6	1.12	1.7	0.94	1.39	1.89

Query executed successfully. | localhost (12.0 RTM) | sa (57) | weather | 00:00:00 | 147 rows