SPIDERFINANCIAL
www.spiderfinancial.com

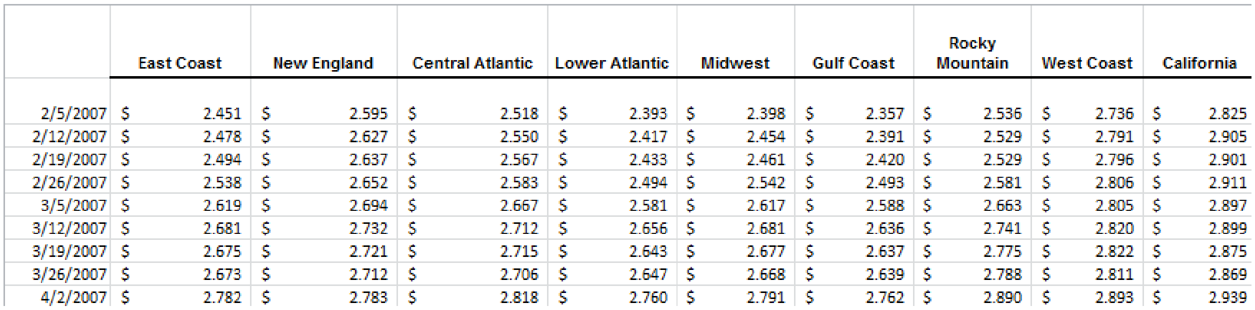
Phone: 1-888-427-9486
1-312-324-0367
Fax: 1-312-238-9092
info@spiderfinancial.com

Making Sense of Diesel Prices

In this case study, we examine closely the highway retail price ($/Gallon) for “No.2 Ultra Low Sulfur (0-15 ppm) Diesel” in the EIA nine (9) PADD regions. We carry on principal component analysis in an attempt to find a minimal subset of the principal components that capture (or explains) the variation in prices between regions with a minimal loss of information.

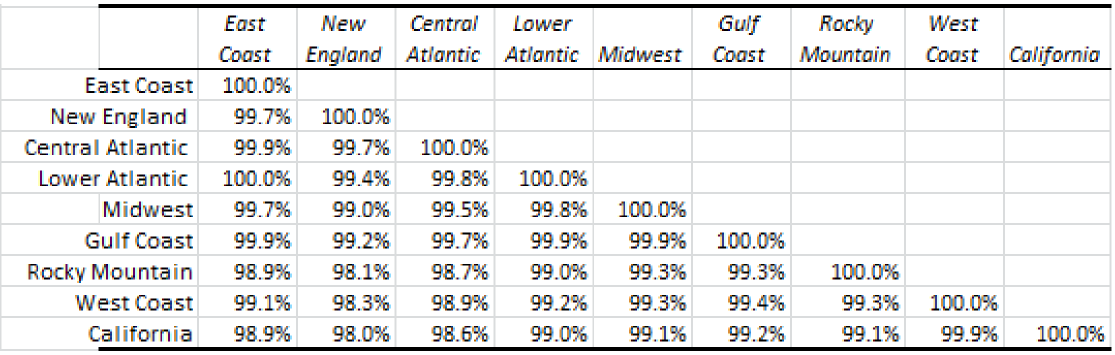
Next, we will use the weekly average highway retail price ($/Gallon) for “No.2 Ultra Low Sulfur (0-15 ppm) Diesel” in the EIA nine (9) PADD regions. The sample data start on February 5th, 2007 and end on May 6th, 2013 (327 observations). Each observation represents the weekly average prices in nine regions:

1. East Coast
2. New England
3. Central Atlantic (PADD 1B)
4. Lower Atlantic (PADD 1C)
5. Midwest
6. Gulf Coast
7. Rocky Mountain
8. West Coast
9. California

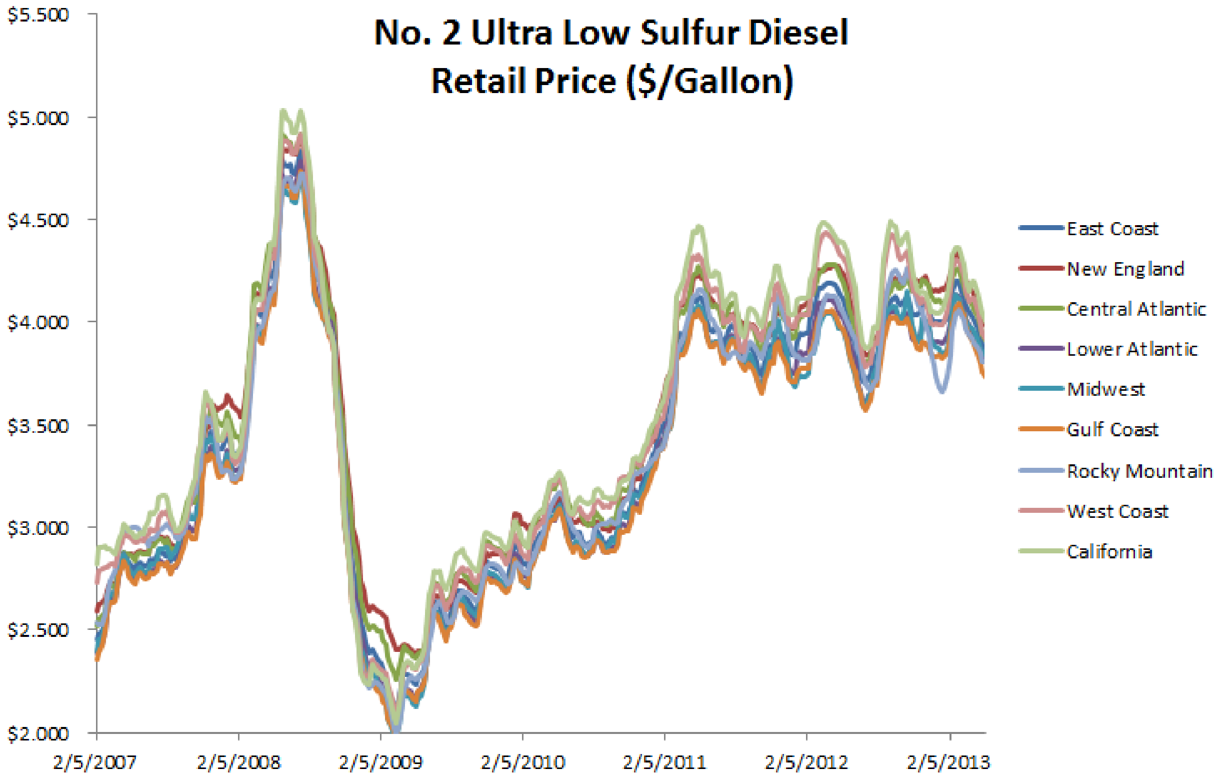


	East Coast	New England	Central Atlantic	Lower Atlantic	Midwest	Gulf Coast	Rocky Mountain	West Coast	California
2/5/2007	$ 2.451	$ 2.595	$ 2.518	$ 2.393	$ 2.398	$ 2.357	$ 2.536	$ 2.736	$ 2.825
2/12/2007	$ 2.478	$ 2.627	$ 2.550	$ 2.417	$ 2.454	$ 2.391	$ 2.529	$ 2.791	$ 2.905
2/19/2007	$ 2.494	$ 2.637	$ 2.567	$ 2.433	$ 2.461	$ 2.420	$ 2.529	$ 2.796	$ 2.901
2/26/2007	$ 2.538	$ 2.652	$ 2.583	$ 2.494	$ 2.542	$ 2.493	$ 2.581	$ 2.806	$ 2.911
3/5/2007	$ 2.619	$ 2.694	$ 2.667	$ 2.581	$ 2.617	$ 2.588	$ 2.663	$ 2.805	$ 2.897
3/12/2007	$ 2.681	$ 2.732	$ 2.712	$ 2.656	$ 2.681	$ 2.636	$ 2.741	$ 2.820	$ 2.899
3/19/2007	$ 2.675	$ 2.721	$ 2.715	$ 2.643	$ 2.677	$ 2.637	$ 2.775	$ 2.822	$ 2.875
3/26/2007	$ 2.673	$ 2.712	$ 2.706	$ 2.647	$ 2.668	$ 2.639	$ 2.788	$ 2.811	$ 2.869
4/2/2007	$ 2.782	$ 2.783	$ 2.818	$ 2.760	$ 2.791	$ 2.762	$ 2.890	$ 2.893	$ 2.939

The weekly prices across the different regions are highly correlated:



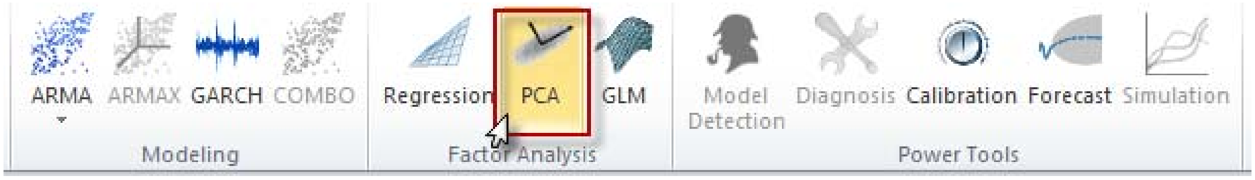
	East Coast	New England	Central Atlantic	Lower Atlantic	Midwest	Gulf Coast	Rocky Mountain	West Coast	California
East Coast	100.0%								
New England	99.7%	100.0%							
Central Atlantic	99.9%	99.7%	100.0%						
Lower Atlantic	100.0%	99.4%	99.8%	100.0%					
Midwest	99.7%	99.0%	99.5%	99.8%	100.0%				
Gulf Coast	99.9%	99.2%	99.7%	99.9%	99.9%	100.0%			
Rocky Mountain	98.9%	98.1%	98.7%	99.0%	99.3%	99.3%	100.0%		
West Coast	99.1%	98.3%	98.9%	99.2%	99.3%	99.4%	99.3%	100.0%	
California	98.9%	98.0%	98.6%	99.0%	99.1%	99.2%	99.1%	99.9%	100.0%



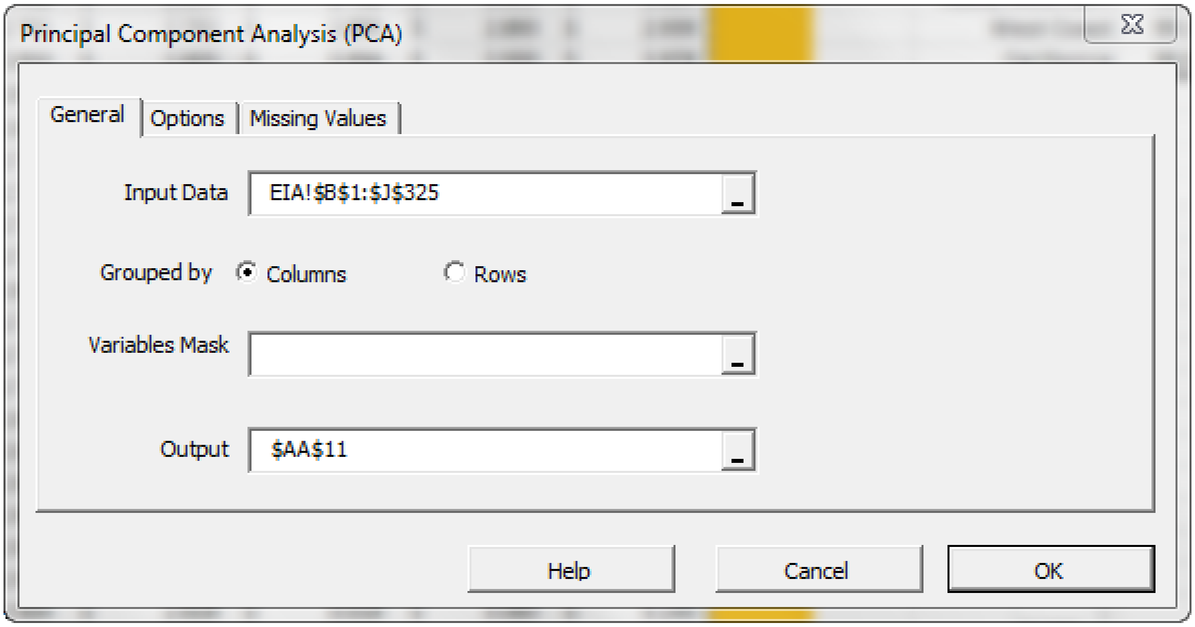
In this tutorial, we'll investigate the drivers behind the price variation among the different regions, and attempt to imply the physical representation of those components using the each region's price loadings.

Process

Now we are ready to conduct our principal component analysis. First, select an empty cell in your worksheet where you wish the output to be generated, then locate and click on the principal component (PCA) icon in the NumXL tab (or toolbar).



The principal component analysis wizard appears.



Principal Component Analysis (PCA)

General | Options | Missing Values

Input Data: EIA!$B$1:$J$325

Grouped by: ● Columns ○ Rows

Variables Mask:

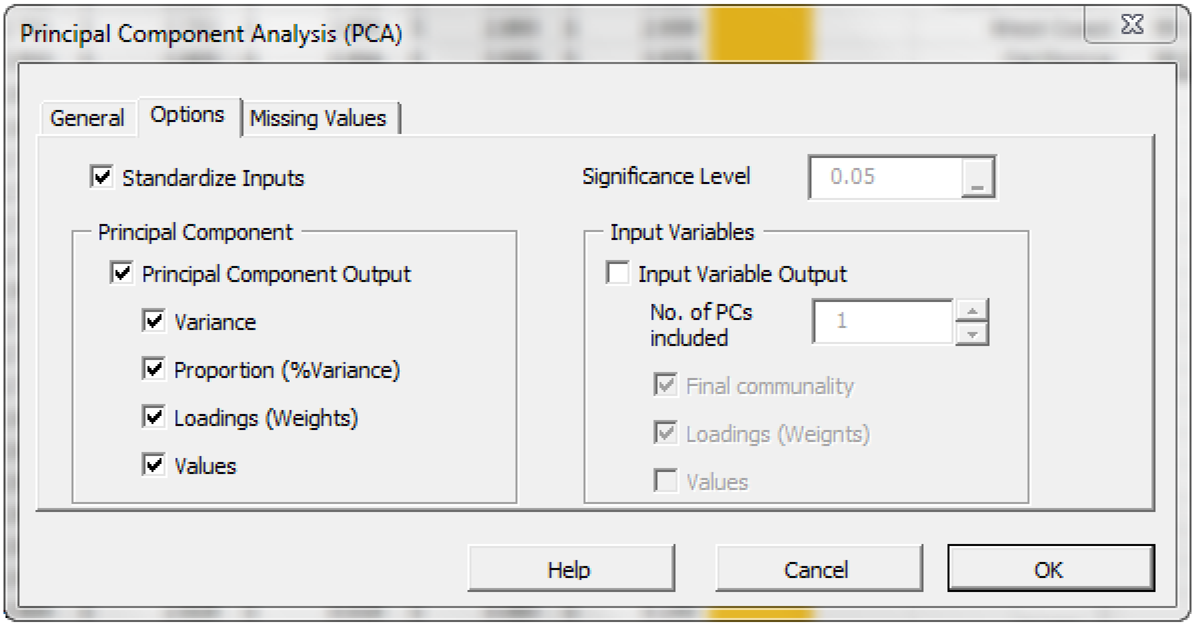
Output: $AA$11

Select the cells range for the five input variable values.

Notes:

1. Leave out the last three observations, so our input data ends on April 15th, 2013. The remaining three points will be used for comparing the forecast values later on.
2. Leave the “Variables Mask” field blank for now. We will revisit this field in later entries.

Next, select the “Options” tab.



Principal Component Analysis (PCA)

General | Options | Missing Values

☑ Standardize Inputs

Significance Level: 0.05

Principal Component

☑ Principal Component Output

☑ Variance

☑ Proportion (%Variance)

☑ Loadings (Weights)

☑ Values

Input Variables

☐ Input Variable Output

No. of PCs included: 1

☑ Final communality

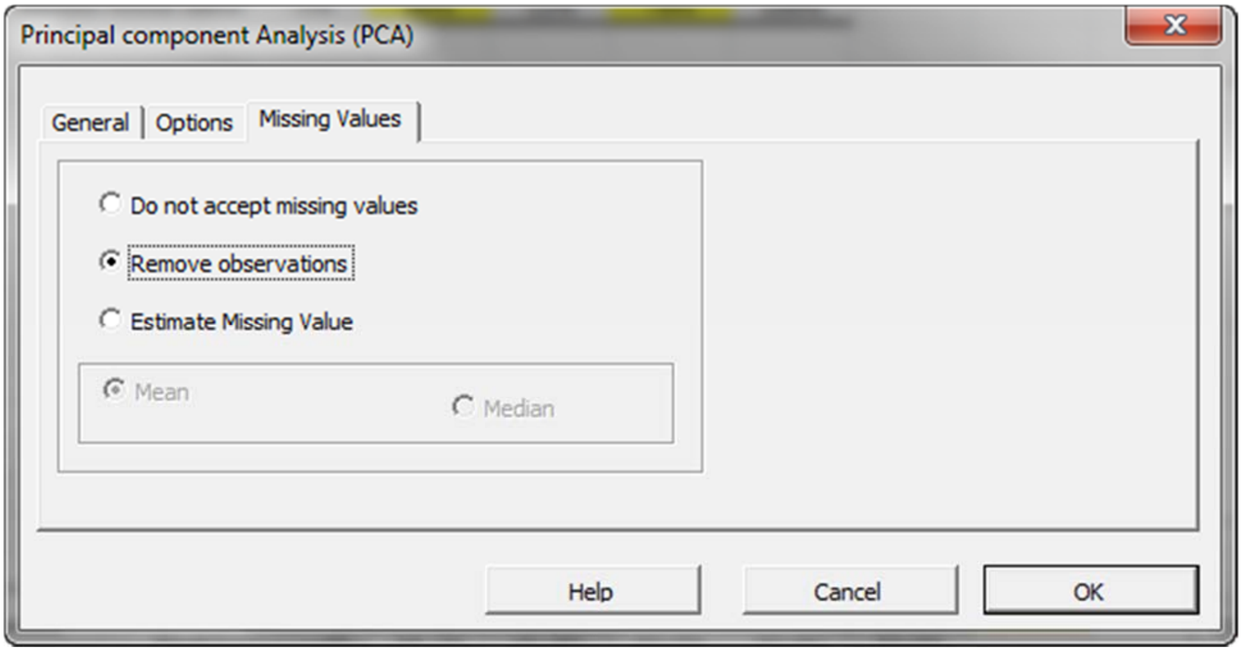
☑ Loadings (Weignts)

☐ Values

Initially, the tab is set to the following values:

- "Standardize Inputs" is checked. **Leave this option checked.**
- "Principal Component Output" is checked. **Leave it checked.**
- The significance level (aka. α) is set to 5%.
- Under "principal component," check the "Values" option, so the generated output tables include the principal component values for different dates.
- "Input Variables" is unchecked. Leave it unchecked**.**

Now, click the "Missing Values" tab.



Principal component Analysis (PCA)

○ Do not accept missing values

● Remove observations

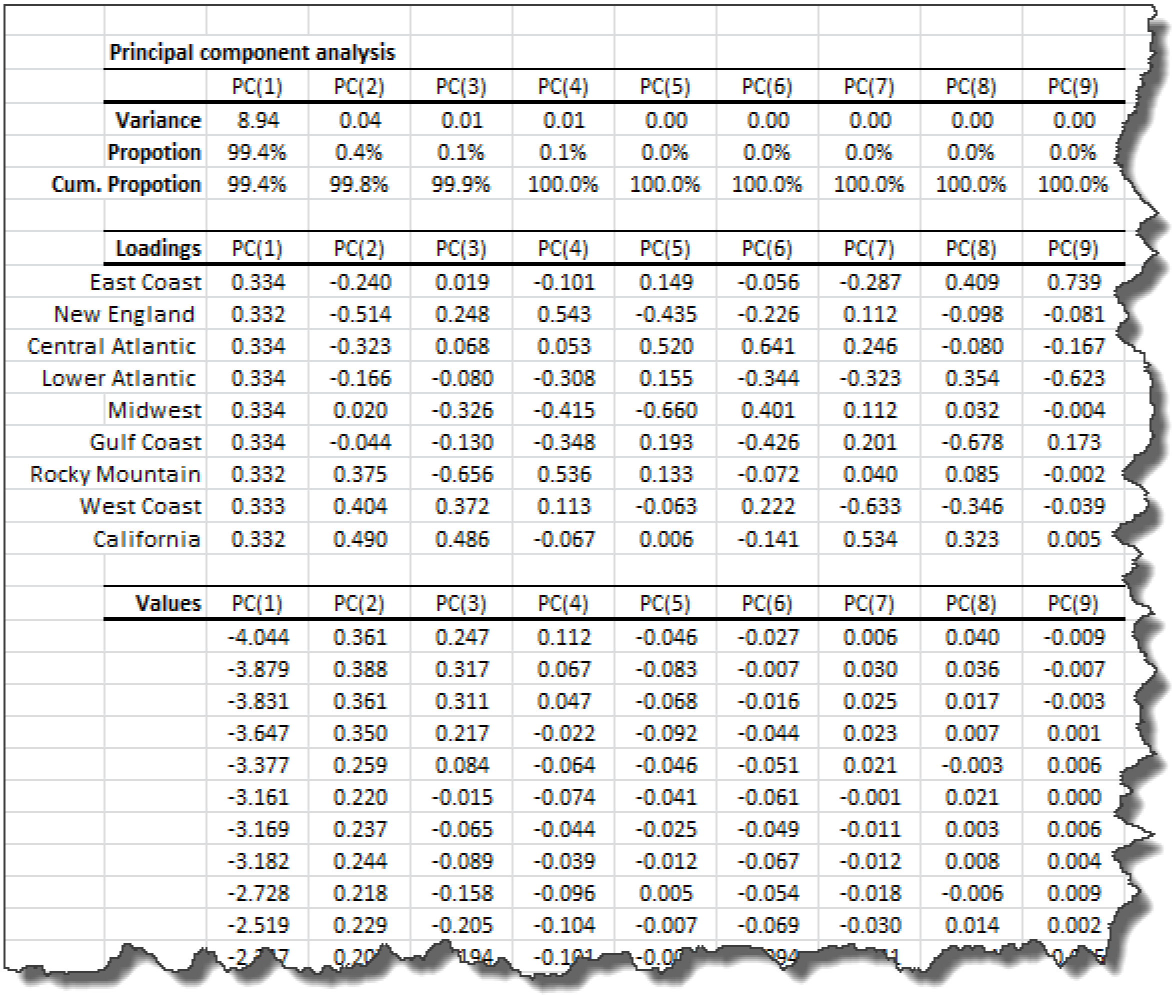
○ Estimate Missing Value

● Mean ○ Median

In this tab, you can select an approach to handle missing values in the data set (X). By default, any missing value found in any observation would exclude the observation from the analysis.

This treatment is a good approach for our analysis, so let's leave it unchanged.

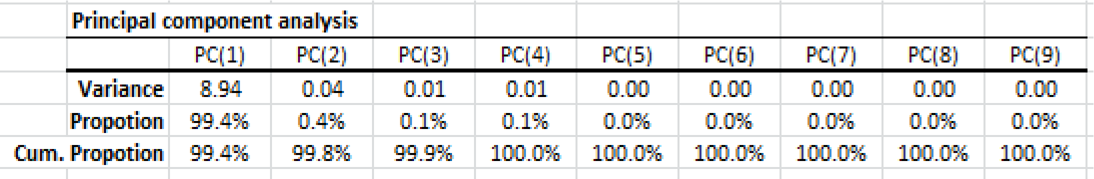
Now, click "OK" to generate the output tables.



Principal component analysis									
	PC(1)	PC(2)	PC(3)	PC(4)	PC(5)	PC(6)	PC(7)	PC(8)	PC(9)
Variance	8.94	0.04	0.01	0.01	0.00	0.00	0.00	0.00	0.00
Propotion	99.4%	0.4%	0.1%	0.1%	0.0%	0.0%	0.0%	0.0%	0.0%
Cum. Propotion	99.4%	99.8%	99.9%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
Loadings	PC(1)	PC(2)	PC(3)	PC(4)	PC(5)	PC(6)	PC(7)	PC(8)	PC(9)
East Coast	0.334	-0.240	0.019	-0.101	0.149	-0.056	-0.287	0.409	0.739
New England	0.332	-0.514	0.248	0.543	-0.435	-0.226	0.112	-0.098	-0.081
Central Atlantic	0.334	-0.323	0.068	0.053	0.520	0.641	0.246	-0.080	-0.167
Lower Atlantic	0.334	-0.166	-0.080	-0.308	0.155	-0.344	-0.323	0.354	-0.623
Midwest	0.334	0.020	-0.326	-0.415	-0.660	0.401	0.112	0.032	-0.004
Gulf Coast	0.334	-0.044	-0.130	-0.348	0.193	-0.426	0.201	-0.678	0.173
Rocky Mountain	0.332	0.375	-0.656	0.536	0.133	-0.072	0.040	0.085	-0.002
West Coast	0.333	0.404	0.372	0.113	-0.063	0.222	-0.633	-0.346	-0.039
California	0.332	0.490	0.486	-0.067	0.006	-0.141	0.534	0.323	0.005
Values	PC(1)	PC(2)	PC(3)	PC(4)	PC(5)	PC(6)	PC(7)	PC(8)	PC(9)
	-4.044	0.361	0.247	0.112	-0.046	-0.027	0.006	0.040	-0.009
	-3.879	0.388	0.317	0.067	-0.083	-0.007	0.030	0.036	-0.007
	-3.831	0.361	0.311	0.047	-0.068	-0.016	0.025	0.017	-0.003
	-3.647	0.350	0.217	-0.022	-0.092	-0.044	0.023	0.007	0.001
	-3.377	0.259	0.084	-0.064	-0.046	-0.051	0.021	-0.003	0.006
	-3.161	0.220	-0.015	-0.074	-0.041	-0.061	-0.001	0.021	0.000
	-3.169	0.237	-0.065	-0.044	-0.025	-0.049	-0.011	0.003	0.006
	-3.182	0.244	-0.089	-0.039	-0.012	-0.067	-0.012	0.008	0.004
	-2.728	0.218	-0.158	-0.096	0.005	-0.054	-0.018	-0.006	0.009
	-2.519	0.229	-0.205	-0.104	-0.007	-0.069	-0.030	0.014	0.002
	-2.[illegible]7	0.20[illegible]	[illegible]194	-0.101	-0.0[illegible]	[illegible]094	[illegible]1	[illegible]	0.[illegible]5

Analysis

1. Statistics

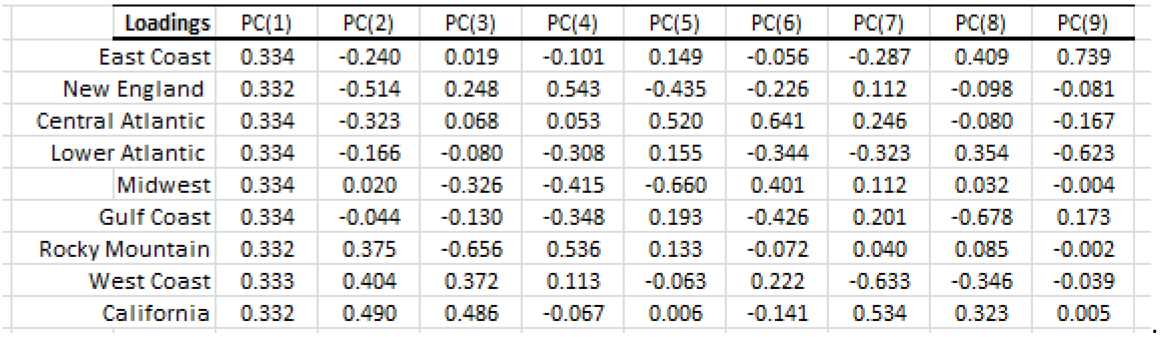


Principal component analysis									
	PC(1)	PC(2)	PC(3)	PC(4)	PC(5)	PC(6)	PC(7)	PC(8)	PC(9)
Variance	8.94	0.04	0.01	0.01	0.00	0.00	0.00	0.00	0.00
Propotion	99.4%	0.4%	0.1%	0.1%	0.0%	0.0%	0.0%	0.0%	0.0%
Cum. Propotion	99.4%	99.8%	99.9%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%

In the table above, we show the variance of each principal component and the proportion of the input (standardized) data set's total variance variable that it accounts for. Examining the table closer, the 1st two components capture 99.8% of the data set variation.

2. Loadings

In the loadings table, we outline the weights of the principal component in each input variable:



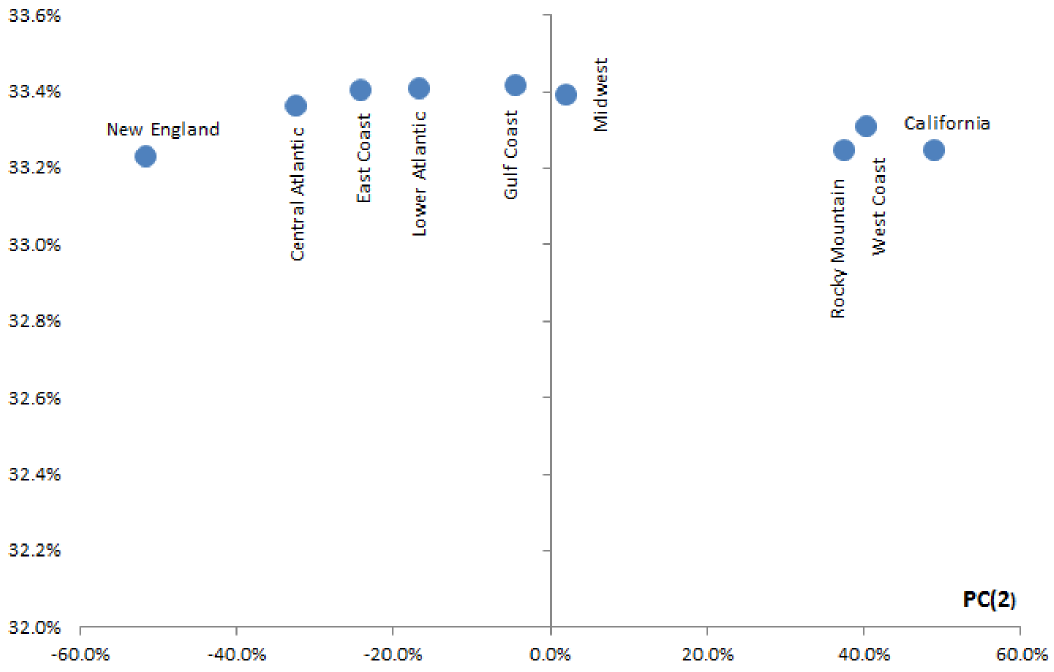
Loadings	PC(1)	PC(2)	PC(3)	PC(4)	PC(5)	PC(6)	PC(7)	PC(8)	PC(9)
East Coast	0.334	-0.240	0.019	-0.101	0.149	-0.056	-0.287	0.409	0.739
New England	0.332	-0.514	0.248	0.543	-0.435	-0.226	0.112	-0.098	-0.081
Central Atlantic	0.334	-0.323	0.068	0.053	0.520	0.641	0.246	-0.080	-0.167
Lower Atlantic	0.334	-0.166	-0.080	-0.308	0.155	-0.344	-0.323	0.354	-0.623
Midwest	0.334	0.020	-0.326	-0.415	-0.660	0.401	0.112	0.032	-0.004
Gulf Coast	0.334	-0.044	-0.130	-0.348	0.193	-0.426	0.201	-0.678	0.173
Rocky Mountain	0.332	0.375	-0.656	0.536	0.133	-0.072	0.040	0.085	-0.002
West Coast	0.333	0.404	0.372	0.113	-0.063	0.222	-0.633	-0.346	-0.039
California	0.332	0.490	0.486	-0.067	0.006	-0.141	0.534	0.323	0.005

Examining the input variables (i.e. region price) loadings for the first component shows a uniform loading for all variables. This can be interpreted as the level-factor (price that is locale-neutral).

For the second factor, the picture is a bit different:

- For all PADD regions in the east, the loading is negative.
- For PADD regions in the west, the loading is positive.
- Gulf coast's loading is slightly negative.
- Midwest loading is slightly positive.

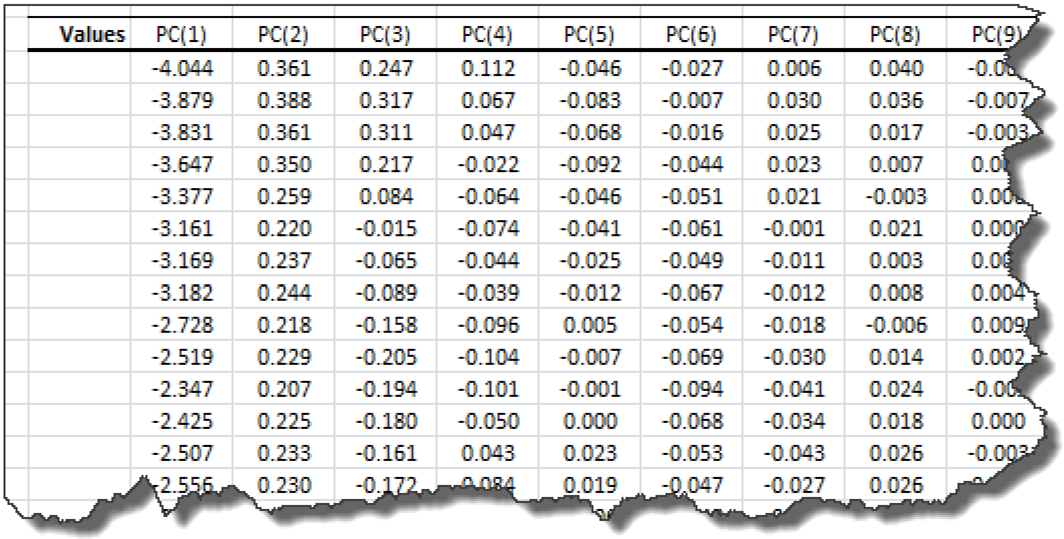
This factor's loading can be viewed as proximity to (or availability of) refinery capacity or import ports (example: New York harbor). The second factor reflects the cost of transportation and tax of the fuel.



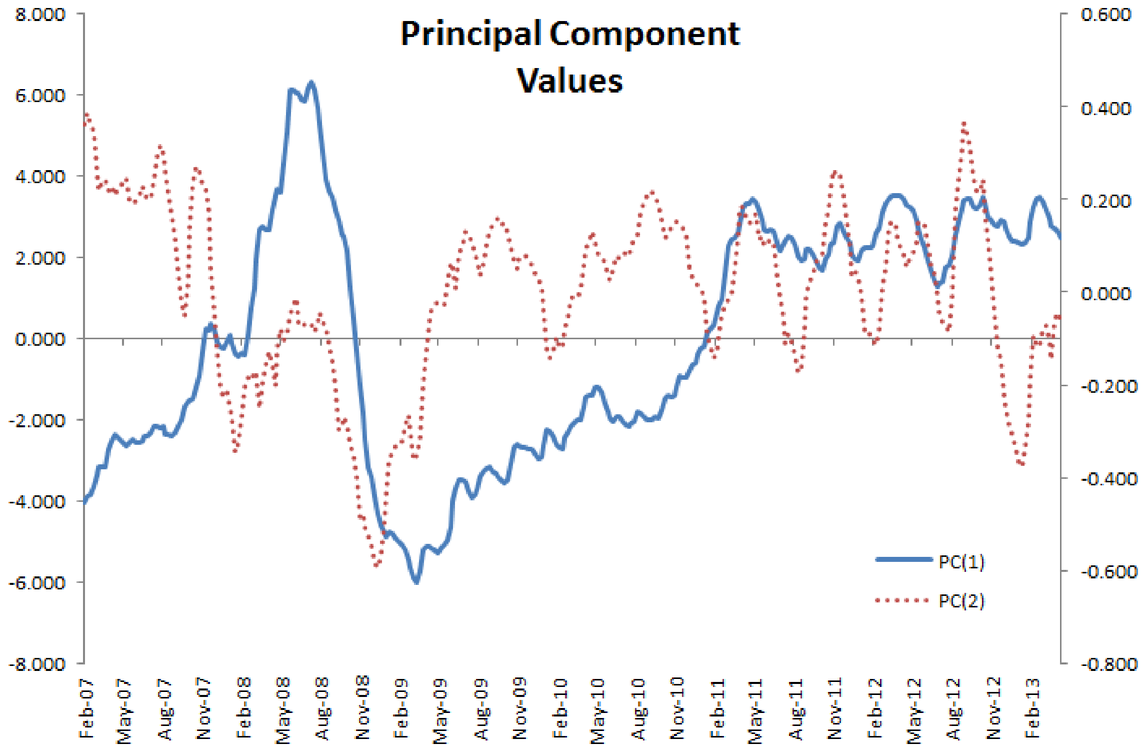
Note: The loadings of the input variables for the 1st component are very comparable, so, in effect, the second component (factor) is what drives the price differential between the different PADD regions.

3. Principal Component Values

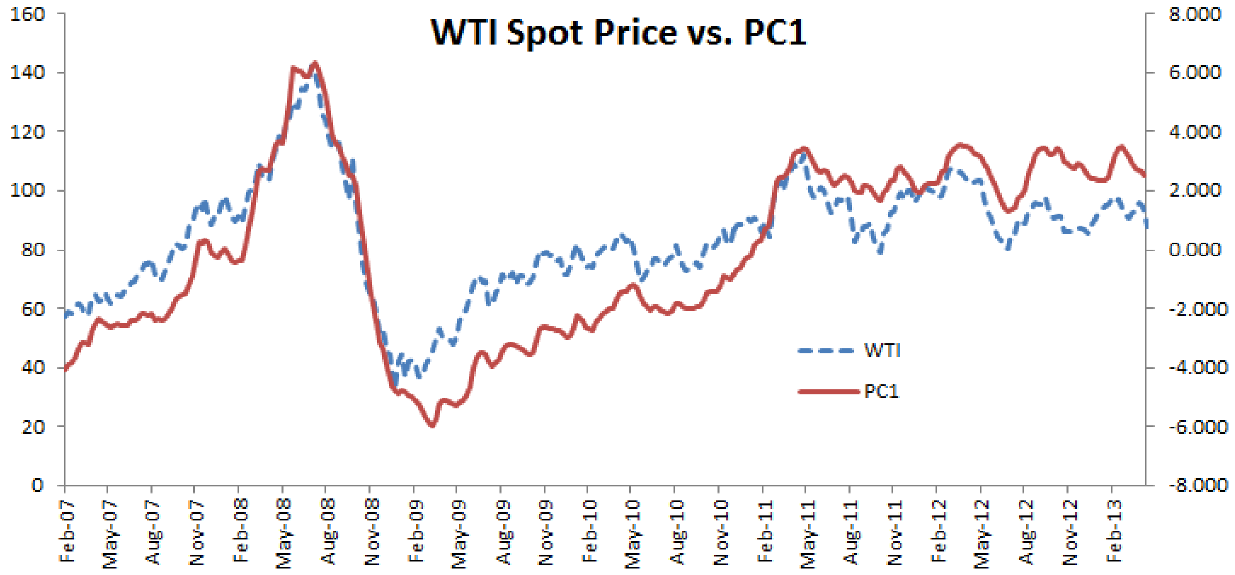
Let's examine the values of the first two principal components.



Values	PC(1)	PC(2)	PC(3)	PC(4)	PC(5)	PC(6)	PC(7)	PC(8)	PC(9)
	-4.044	0.361	0.247	0.112	-0.046	-0.027	0.006	0.040	-0.00
	-3.879	0.388	0.317	0.067	-0.083	-0.007	0.030	0.036	-0.007
	-3.831	0.361	0.311	0.047	-0.068	-0.016	0.025	0.017	-0.003
	-3.647	0.350	0.217	-0.022	-0.092	-0.044	0.023	0.007	0.00
	-3.377	0.259	0.084	-0.064	-0.046	-0.051	0.021	-0.003	0.00
	-3.161	0.220	-0.015	-0.074	-0.041	-0.061	-0.001	0.021	0.000
	-3.169	0.237	-0.065	-0.044	-0.025	-0.049	-0.011	0.003	0.00
	-3.182	0.244	-0.089	-0.039	-0.012	-0.067	-0.012	0.008	0.004
	-2.728	0.218	-0.158	-0.096	0.005	-0.054	-0.018	-0.006	0.009
	-2.519	0.229	-0.205	-0.104	-0.007	-0.069	-0.030	0.014	0.002
	-2.347	0.207	-0.194	-0.101	-0.001	-0.094	-0.041	0.024	-0.00
	-2.425	0.225	-0.180	-0.050	0.000	-0.068	-0.034	0.018	0.000
	-2.507	0.233	-0.161	0.043	0.023	-0.053	-0.043	0.026	-0.003
	-2.556	0.230	-0.172	0.084	0.019	-0.047	-0.027	0.026	[illegible]



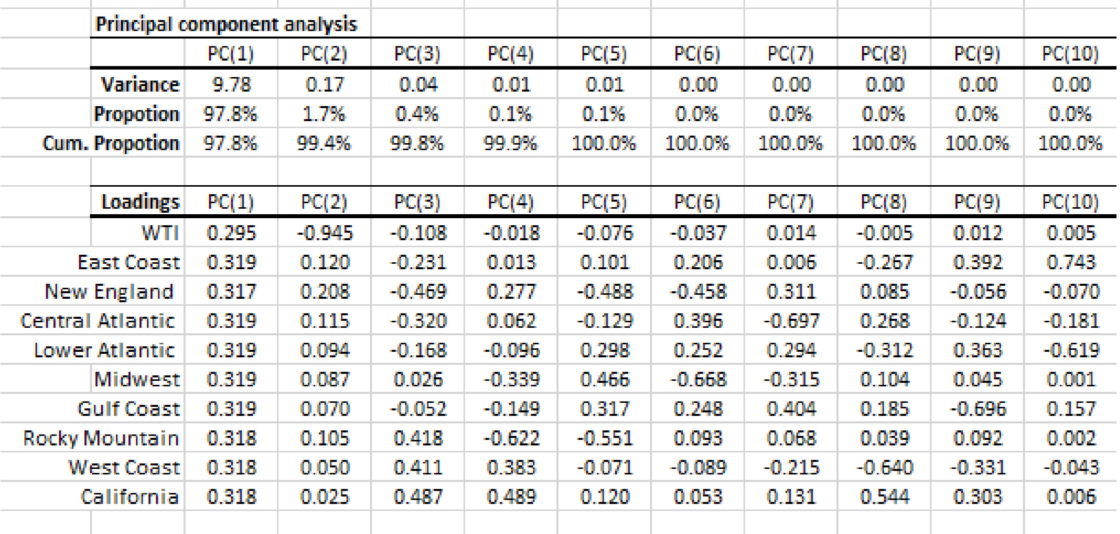
The two time series exhibit some seasonality, although it is more apparent in the second factor. Furthermore, the first principal component exhibits a pattern close to crude oil prices.



NOTE: The lack of an exact match may be attributed to other costs incurred in the making of No. 2 Ultra Low Sulfur Diesel: labor, energy prices, raw material, etc. Furthermore, refineries build up inventory of products (e.g. diesel) in anticipation of the seasonal demand peaks, so there may be a lag.

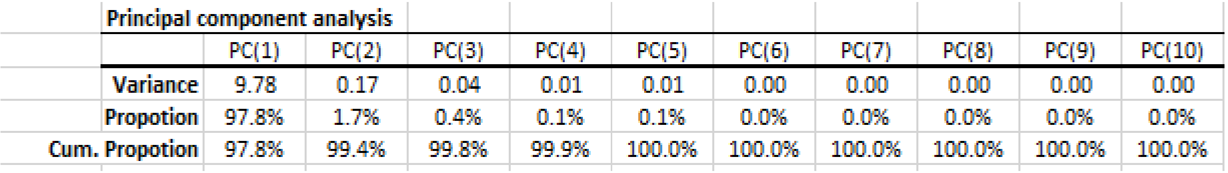
4. Adding WTI Spot Prices

As a last thought on the WTI spot price, we will include the WTI spot price in our input data set and re-examine the input loadings. Intuitively, adding an input variable for raw material price (i.e. crude oil) with the finished products prices in the same data set will likely reveal another driver: cost of production.



Principal component analysis										
	PC(1)	PC(2)	PC(3)	PC(4)	PC(5)	PC(6)	PC(7)	PC(8)	PC(9)	PC(10)
Variance	9.78	0.17	0.04	0.01	0.01	0.00	0.00	0.00	0.00	0.00
Propotion	97.8%	1.7%	0.4%	0.1%	0.1%	0.0%	0.0%	0.0%	0.0%	0.0%
Cum. Propotion	97.8%	99.4%	99.8%	99.9%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
Loadings	PC(1)	PC(2)	PC(3)	PC(4)	PC(5)	PC(6)	PC(7)	PC(8)	PC(9)	PC(10)
WTI	0.295	-0.945	-0.108	-0.018	-0.076	-0.037	0.014	-0.005	0.012	0.005
East Coast	0.319	0.120	-0.231	0.013	0.101	0.206	0.006	-0.267	0.392	0.743
New England	0.317	0.208	-0.469	0.277	-0.488	-0.458	0.311	0.085	-0.056	-0.070
Central Atlantic	0.319	0.115	-0.320	0.062	-0.129	0.396	-0.697	0.268	-0.124	-0.181
Lower Atlantic	0.319	0.094	-0.168	-0.096	0.298	0.252	0.294	-0.312	0.363	-0.619
Midwest	0.319	0.087	0.026	-0.339	0.466	-0.668	-0.315	0.104	0.045	0.001
Gulf Coast	0.319	0.070	-0.052	-0.149	0.317	0.248	0.404	0.185	-0.696	0.157
Rocky Mountain	0.318	0.105	0.418	-0.622	-0.551	0.093	0.068	0.039	0.092	0.002
West Coast	0.318	0.050	0.411	0.383	-0.071	-0.089	-0.215	-0.640	-0.331	-0.043
California	0.318	0.025	0.487	0.489	0.120	0.053	0.131	0.544	0.303	0.006

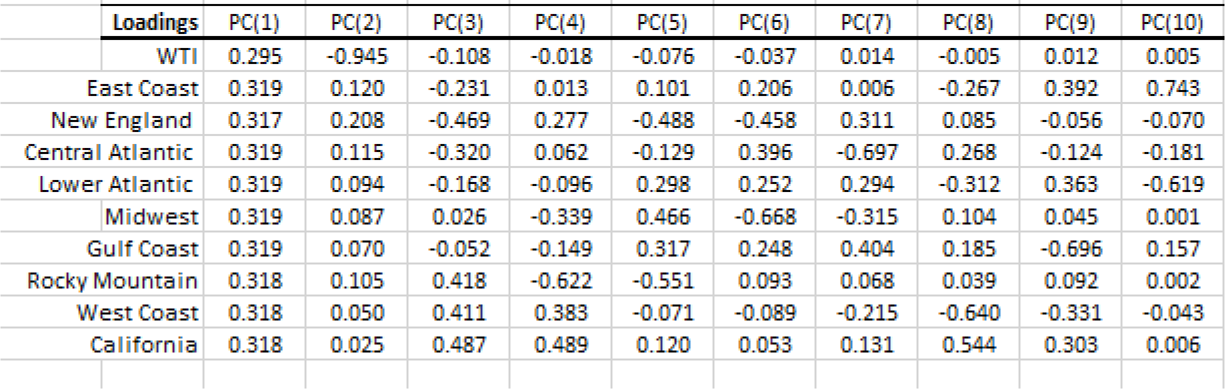
Now, we'd need three drivers to account for 99.8% of the price variation.



	Principal component analysis									
	PC(1)	PC(2)	PC(3)	PC(4)	PC(5)	PC(6)	PC(7)	PC(8)	PC(9)	PC(10)
Variance	9.78	0.17	0.04	0.01	0.01	0.00	0.00	0.00	0.00	0.00
Propotion	97.8%	1.7%	0.4%	0.1%	0.1%	0.0%	0.0%	0.0%	0.0%	0.0%
Cum. Propotion	97.8%	99.4%	99.8%	99.9%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%

IMPORTANT: The principal components of the new data set **are not necessarily the same** as the ones we computed earlier with only the diesel prices.

Let's look at the input variable's loadings in each principal component (i.e. driver):

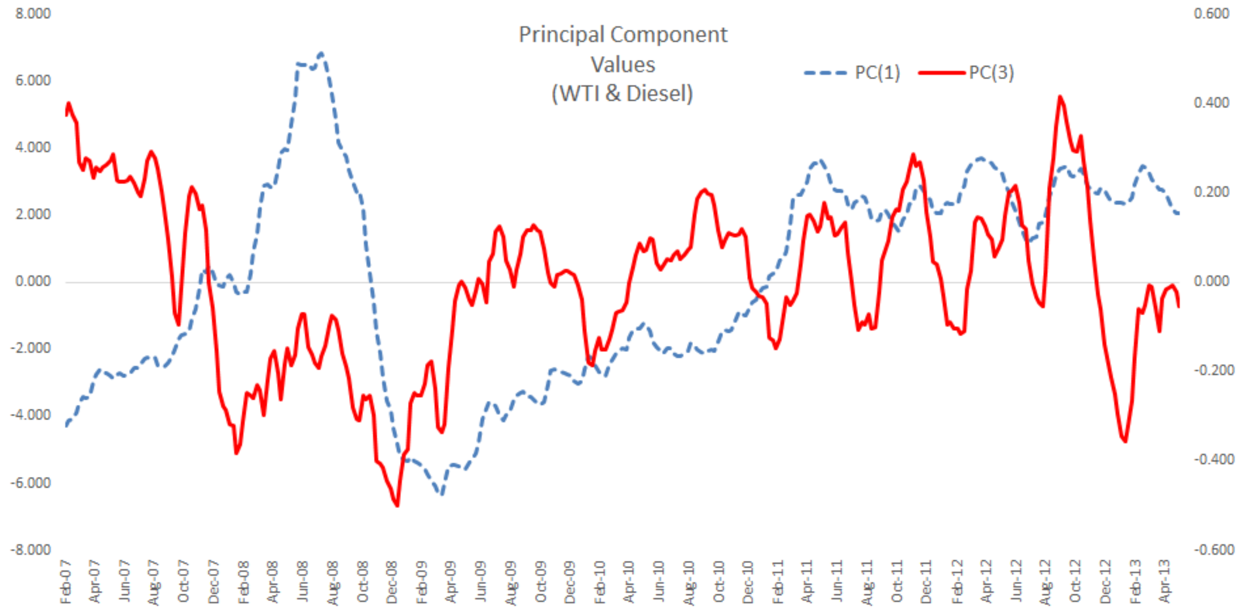


Loadings	PC(1)	PC(2)	PC(3)	PC(4)	PC(5)	PC(6)	PC(7)	PC(8)	PC(9)	PC(10)
WTI	0.295	-0.945	-0.108	-0.018	-0.076	-0.037	0.014	-0.005	0.012	0.005
East Coast	0.319	0.120	-0.231	0.013	0.101	0.206	0.006	-0.267	0.392	0.743
New England	0.317	0.208	-0.469	0.277	-0.488	-0.458	0.311	0.085	-0.056	-0.070
Central Atlantic	0.319	0.115	-0.320	0.062	-0.129	0.396	-0.697	0.268	-0.124	-0.181
Lower Atlantic	0.319	0.094	-0.168	-0.096	0.298	0.252	0.294	-0.312	0.363	-0.619
Midwest	0.319	0.087	0.026	-0.339	0.466	-0.668	-0.315	0.104	0.045	0.001
Gulf Coast	0.319	0.070	-0.052	-0.149	0.317	0.248	0.404	0.185	-0.696	0.157
Rocky Mountain	0.318	0.105	0.418	-0.622	-0.551	0.093	0.068	0.039	0.092	0.002
West Coast	0.318	0.050	0.411	0.383	-0.071	-0.089	-0.215	-0.640	-0.331	-0.043
California	0.318	0.025	0.487	0.489	0.120	0.053	0.131	0.544	0.303	0.006

Notes:

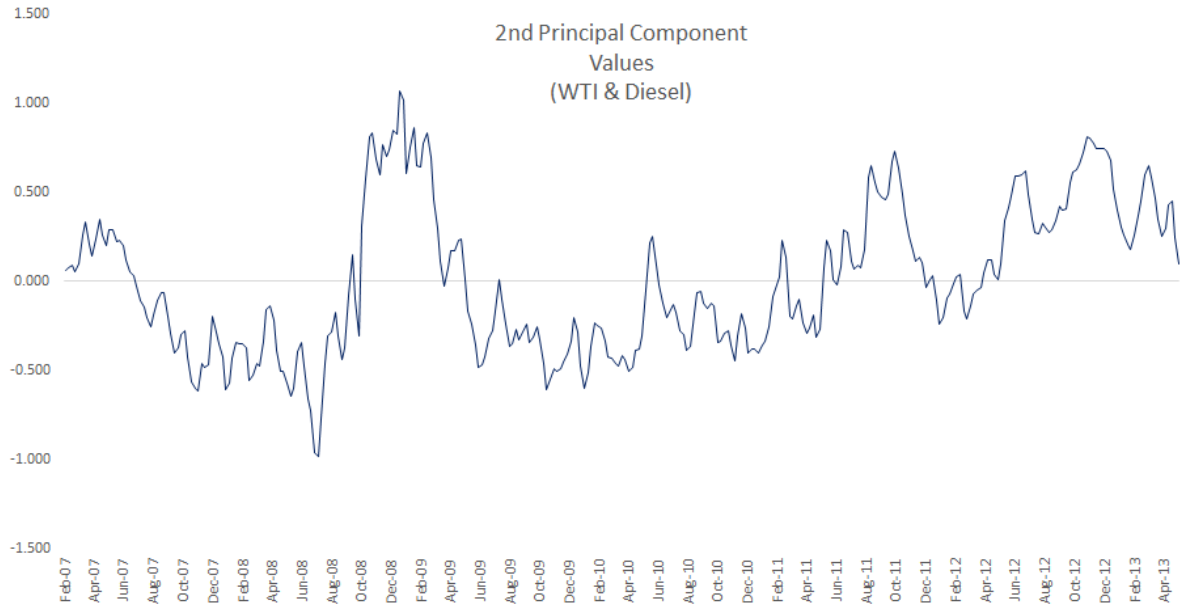
1. The loadings for the first component are similar to ones we calculated earlier with only the diesel prices. Note that the WTI loadings are slightly lower than their diesel counterparts. Again, we'll designate this factor as the general price level (region neutral).
2. The loadings for the second factor are very different now, and the loading for WTI is negative (-%94.5) while all the rest are positive. We can designate this factor as the cost of cracking diesel of crude.
3. The loading for the 3rd factor is very similar to the loading of the second component in the earlier data sets. Again, the loading varies based on the location (east vs. west). The WTI, Gulf Coast, and Midwest are almost neutral.

Let's now plot the factor's time series.



In general terms, the first and the third component are very similar to the first two components we calculated earlier (w/o WTI in the data set).

For the second component, we hypothesize this one as the proxy of the diesel cracking cost.



Conclusion

In this tutorial, we examined a non-trivial application for PCA. We attempted to explain the price variation among different PADD regions by uncovering the driving forces behind the prices.

What's next?

We mentioned earlier that refineries build up inventory of products (e.g. diesel) in anticipation of the seasonal demand peaks, so there may be a lag. How do we capture this variable? Futures: we can include the future prices of the diesel and crude oil.

Why?

Intuitively, futures prices reflect market anticipation of (1) future demand (2) future storage cost, and possibly a premium for supply scarcity.

This application is intended to give you a sample of how to apply PCA and time series, as well as how to use or interpret the variables' loadings in deriving a practical proxy for them.

In sum, PCA is a mathematical procedure that NumXL can help you execute. Making sense of and interpreting the results is where your expertise and intuition are indispensable.