Statistical Analysis of the Similarity of Environmental Data Populations in the Aleutian Islands

Performed by: Lynetta Campbell, B.S. Chemical Engineering, M.S. Management Science, M.S. Mathematics with Certification in Statistical Analytics

Executive Summary:
This report addresses the question of whether or not the environmental data sets (drawn from four buoys and three airport) assembled for the Aleutian Islands Risk Assessment appear to be drawing from the same population of weather. A Nuka statistician performed a comparison of the data sets. It was determined that statistically significant differences exist between all data sets, making the practice of using the data from one source to make inferences about another inadvisable.

Review of Theory:
When one is studying a continuous variable drawn from different populations, the test of whether or not the populations are actually the same involves comparing the means and the variances of the samples. If both of these are the same, then the populations are judged to be the same.

In the situation where there are only two populations, then an independent samples t-test can be used to address the null hypothesis that the means for a given variable drawn from each group are equal, and a test such as Levene's or Bartlett's test can answer the question about whether or not the variances are equal. Bartlett's test is generally held to be better if the variables are normally distributed, while Levene's test is more robust against non-normality. If there are more than two populations, then an ANOVA F-test must be used to test the equivalent means hypothesis. Of course, the ANOVA F-test can also be used to test the similarity of two populations, providing the same results as the t-test.

The multivariate analogs of the above are a Hotelling's T-square test of the hypothesis that the mean vectors for multiple variables from two populations are equal or a Wilk's Lambda statistic from a MANOVA type regression if two or more mean vectors are being compared. Box's M test is used to address the question of whether or not the variance-covariance structures of the populations are the same. As in the univariate case, these tests are much more reliable when the distribution of the variables is multivariate normal.

Data Exploration and Analysis

Buoy Data
The entire data set has missing values in it as detailed in Nuka's report *Characterizing Environmental Conditions in the Aleutian Islands*. For the purposes of this study, only observations that were complete (i.e. no missing information) were included. The first point considered involved balance of the samples. It was important to know how often each buoy was sampled.Table 1, below, shows how many times each buoy was sampled with a complete record during each month and year.

Table 1: Sampling Frequency of Marine Buoys					
Month	**Year**	**Number records 46070**	**Number records 46072**	**Number records 46073**	**Number records 46075**
1	2004		693		24
1	2005		48	168	
1	2006	67	517	23	
1	2007	560	48		
1	2008	599	546	24	120
1	2009	53		95	497
1	2010			598	215
1	2011	96	72	537	742
1	2012	648	640		597
2	2004		657		24
2	2005		48	167	
2	2006	72	444	18	
2	2007	329	46		
2	2008	559	512	24	115
2	2009	44		94	103
2	2010			568	215
2	2011	96	72	476	671
2	2012	595	592		546
3	2004		706		24
3	2005		48	168	
3	2006	72	474	24	
3	2007	161	47		
3	2008	602	567	24	105
3	2009	140		94	45
3	2010			504	568
3	2011	96	71	301	742
3	2012	603	601		144
4	2004		671		24
4	2005		48	167	
4	2006	71	64	6	
4	2007	163	43		
4	2008	144	543	22	119
4	2009	153		94	48
4	2010			627	648
4	2011	96	72	95	720
4	2012	617	617		144
5	2004		681		477

5	2005		48	600	
5	2006	70	70		
5	2007	161	45		
5	2008	155	533	20	120
5	2009	205		92	48
5	2010			660	672
5	2011	96	72	72	742
5	2012	641	644		144
6	2004		629		24
6	2005		47	598	
6	2006	69	66		
6	2007	165	42		
6	2008	592	119	211	117
6	2009	147		92	48
6	2010			613	648
6	2011	94	70	69	719
6	2012	617	622		240
7	2004		705		24
7	2005		45	624	
7	2006	72	70		
7	2007	160	45		
7	2008	610	118	236	144
7	2009	225		91	47
7	2010			638	672
7	2011	96	72	63	744
7	2012	639	648		599
8	2004		658		40
8	2005		45	620	
8	2006	64	67		
8	2007	620	47		
8	2008	513	119	12	569
8	2009	122		457	48
8	2010			637	672
8	2011	101	72	48	744
8	2012	632	645		578
9	2003		48		
9	2004		601		48
9	2005		43	594	
9	2006	401	71		
9	2007	593	252		
9	2008	502	114	12	546

9	2009	177		522	47
9	2010			615	647
9	2011	528	416	47	720
9	2012	192	190		120
10	2004		648		48
10	2005		457	614	
10	2006	505	60		
10	2007	591	47		
10	2008	270	119	12	563
10	2009	45		538	48
10	2010			178	672
10	2011	549	526	45	744
10	2012	190	189		120
11	2004		592		44
11	2005		460	547	
11	2006	477	60		
11	2007	585	452		
11	2008	272	119	12	535
11	2009	38		510	48
11	2010			185	647
11	2011	528	501	46	719
11	2012	191	191		118
12	2004		389		24
12	2005		484	614	
12	2006	487	58		
12	2007	597	475		
12	2008	170	120	10	568
12	2009	37		547	48
12	2010			186	672
12	2011	552	521	47	742
12	2012	189	192		96

Obviously, the buoys were not sampled with complete records at the same frequency. Therefore, a series of analyses were decided upon in which only short periods that were populated with data from all buoys were considered.

January 2008

Consider the month of January, 2008, a month for which some data is available from all buoys. Table 2 displays the mean vectors for each buoy. The analysis will focus on the first ten variables.

Table 2: Mean Vectors for Buoy Data							
BuoyID	N Obs	Variable	N	Mean	Std Dev	Min	Max
46070	599	WDIR	599	169.1469	101.0792	0	359
		WSPD_mpers	599	10.40083	5.229887	0.3	28
		GST_mpers	599	13.22638	6.682454	1.1	37.6
		WVHT_m	599	4.500835	2.425135	0.8	14.59
		DPD_sec	599	10.36845	2.459966	3.45	16
		APD_sec	599	7.227763	1.407372	3.84	11.68
		ATMP_degC	599	1.777296	3.58894	-2.6	12.4
		WTMP_degC	599	3.557262	2.908759	1.9	12.8
		PRES_hPa	599	996.8576	13.17166	960.5	1022
		DEWP_degC	599	-0.4778	4.483111	-8.2	11.6
		Day_illumination_hours	599	9.193139	2.499146	7.27	17.27
		Sky_illumination_hours	599	1.402721	0.119871	1.2	1.9
		Total_illumination_hours	599	10.59559	2.555903	8.78	19.18
46072	546	WDIR	546	171.7674	76.06665	0	358
		WSPD_mpers	546	9.940476	4.463026	0.4	23.4
		GST_mpers	546	12.69945	5.358743	1.6	30.4
		WVHT_m	546	4.847271	1.756485	1.02	11.66
		DPD_sec	546	11.64275	2.292507	5.56	17.39
		APD_sec	546	8.056282	1.268567	5.21	11.89
		ATMP_degC	546	2.85696	0.980001	0.1	5.8
		WTMP_degC	546	3.856044	0.298267	2.9	4.8
		PRES_hPa	546	1003.34	13.22211	967.8	1030.5
		DEWP_degC	546	-0.21007	2.140888	-6.6	4.3
		Day_illumination_hours	546	9.195861	1.568549	7.92	14.92
		Sky_illumination_hours	546	1.245843	0.045431	1.12	1.33
		Total_illumination_hours	546	10.44169	1.550028	9.25	16.2
46073	24	WDIR	24	187.875	108.6367	9	347
		WSPD_mpers	24	4.925	2.210695	2.5	9.8
		GST_mpers	24	5.7875	2.674568	2.9	11.7
		WVHT_m	24	0.774167	0.29339	0.47	1.53
		DPD_sec	24	10.69417	4.326587	5.56	17.39
		APD_sec	24	5.721667	0.464652	5.03	6.73
		ATMP_degC	24	8.1	1.389714	6.4	9.7
		WTMP_degC	24	8.2375	1.380867	6.8	9.8
		PRES_hPa	24	1007.94	9.247181	995.4	1018.9
		DEWP_degC	24	5.941667	2.05763	2.5	8.2
		Day_illumination_hours	24	16.55	0.715056	15.85	17.25
		Sky_illumination_hours	24	1.71	0.183871	1.53	1.89
		Total_illumination_hours	24	18.265	0.904034	17.38	19.15

46075	120	WDIR	120	252.2083	68.68548	112	330
		WSPD_mpers	120	8.65	4.313551	1.4	18.3
		GST_mpers	120	10.7075	5.600378	2.1	24.4
		WVHT_m	120	3.057917	2.798955	0.92	10.85
		DPD_sec	120	9.71375	3.591184	4.17	16.67
		APD_sec	120	6.647417	1.894749	4.39	11.54
		ATMP_degC	120	7.940833	2.584618	3.2	12.3
		WTMP_degC	120	8.825	2.341236	5.5	11.3
		PRES_hPa	120	1008.99	4.409896	997.4	1015
		DEWP_degC	120	4.591667	3.333295	-1.3	10
		Day_illumination_hours	120	11.604	2.859137	7.77	15.69
		Sky_illumination_hours	120	1.304	0.11175	1.17	1.46
		Total_illumination_hours	120	12.906	2.878673	9.17	17.15

The means for the various buoys don't look to be the same, but before testing the hypothesis, a bit more data exploration is warranted. To facilitate this, the dimensionality of the dataset was first reduced using principal components analysis. A PCA based on the correlation matrix (i.e. on standardized values of the data) revealed that 80% of the variation contained in this data set can be captured by the first three principal components (see Figure 1, below).

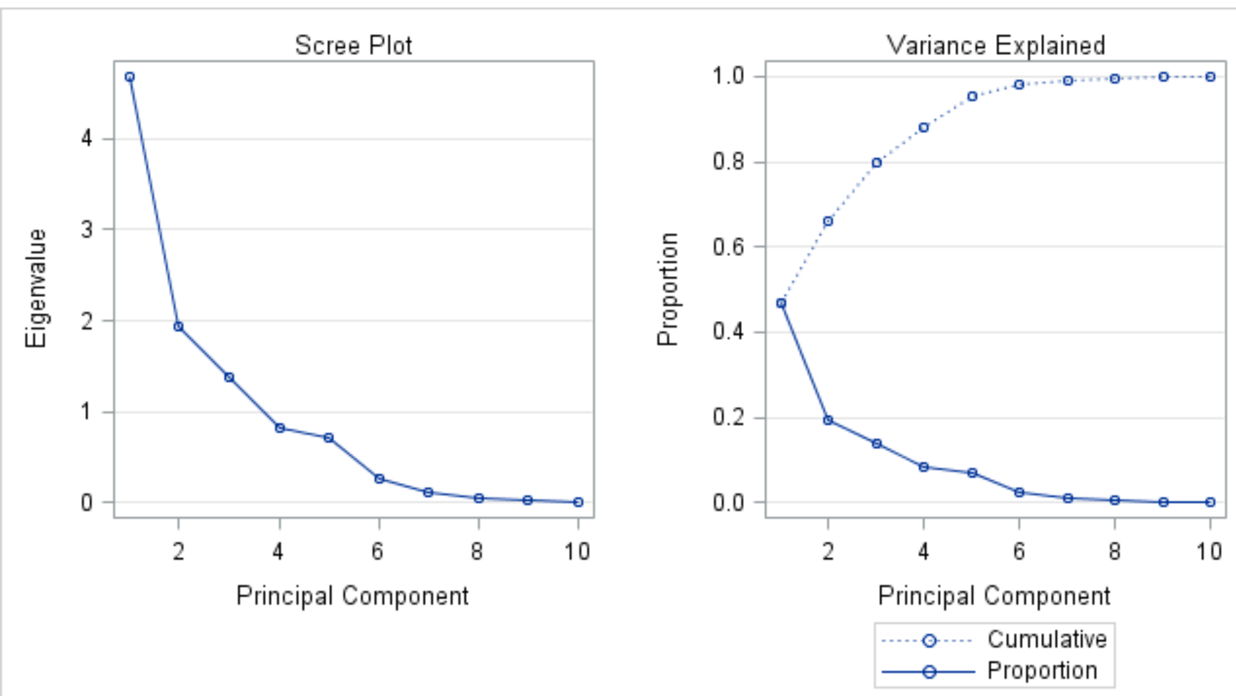

Figure 1: Scree Plot and Variance Explained for the first three Prinicipal Components

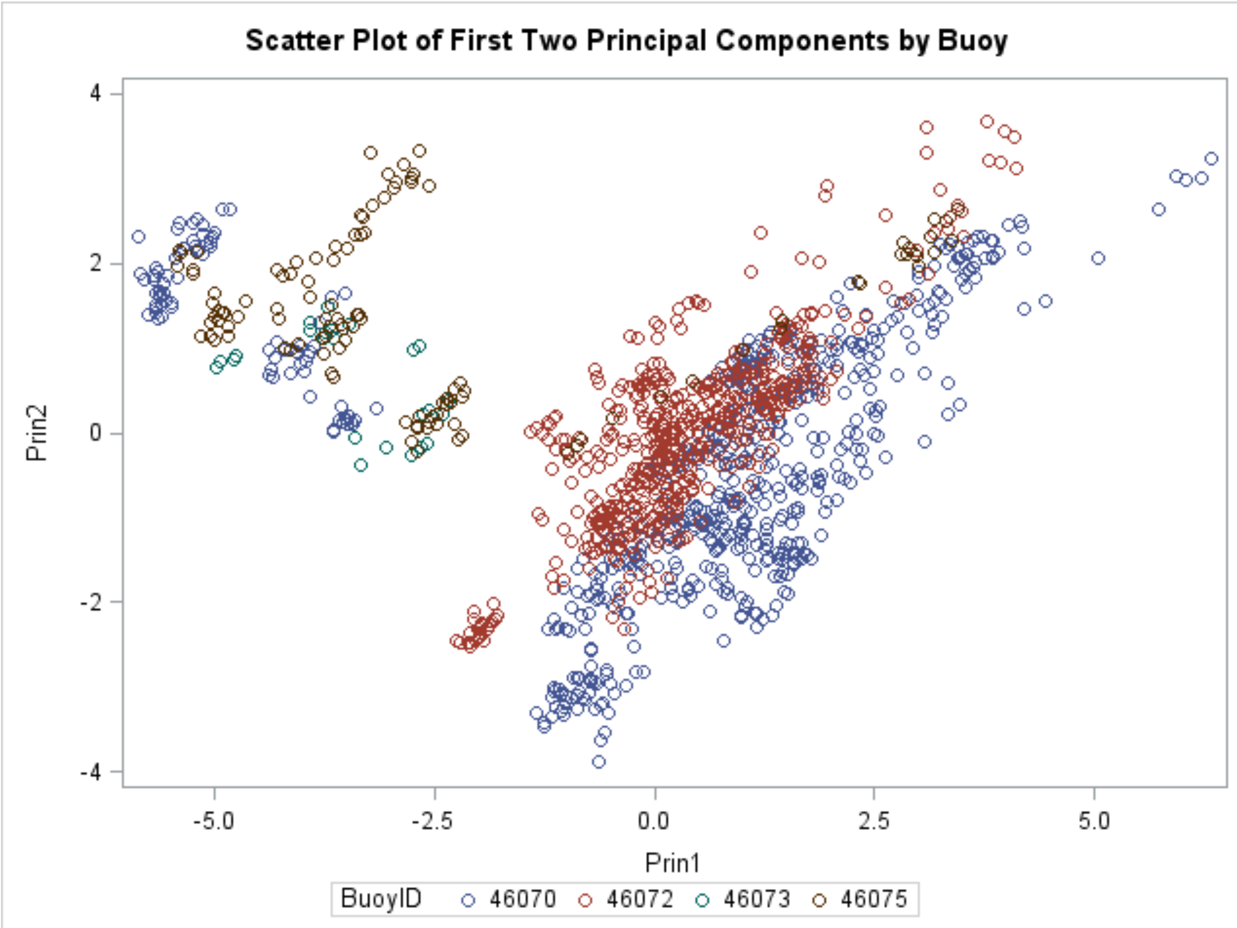

Figure 2: Scatter Plot for the first two Principal Components

Figure 2, with its highly defined clusters, visually suggests significant differences in the buoys. When only the first buoy is examined, it is seen that there are even distinct clusters within this subgroup indicating that weather varies appreciably over the course of a month. The portion of the data for which the values of the first principal component are the most negative correspond to a period from January 7 through January 9 that is categorized by much lower wind speeds and smaller wave heights than during the rest of the month.

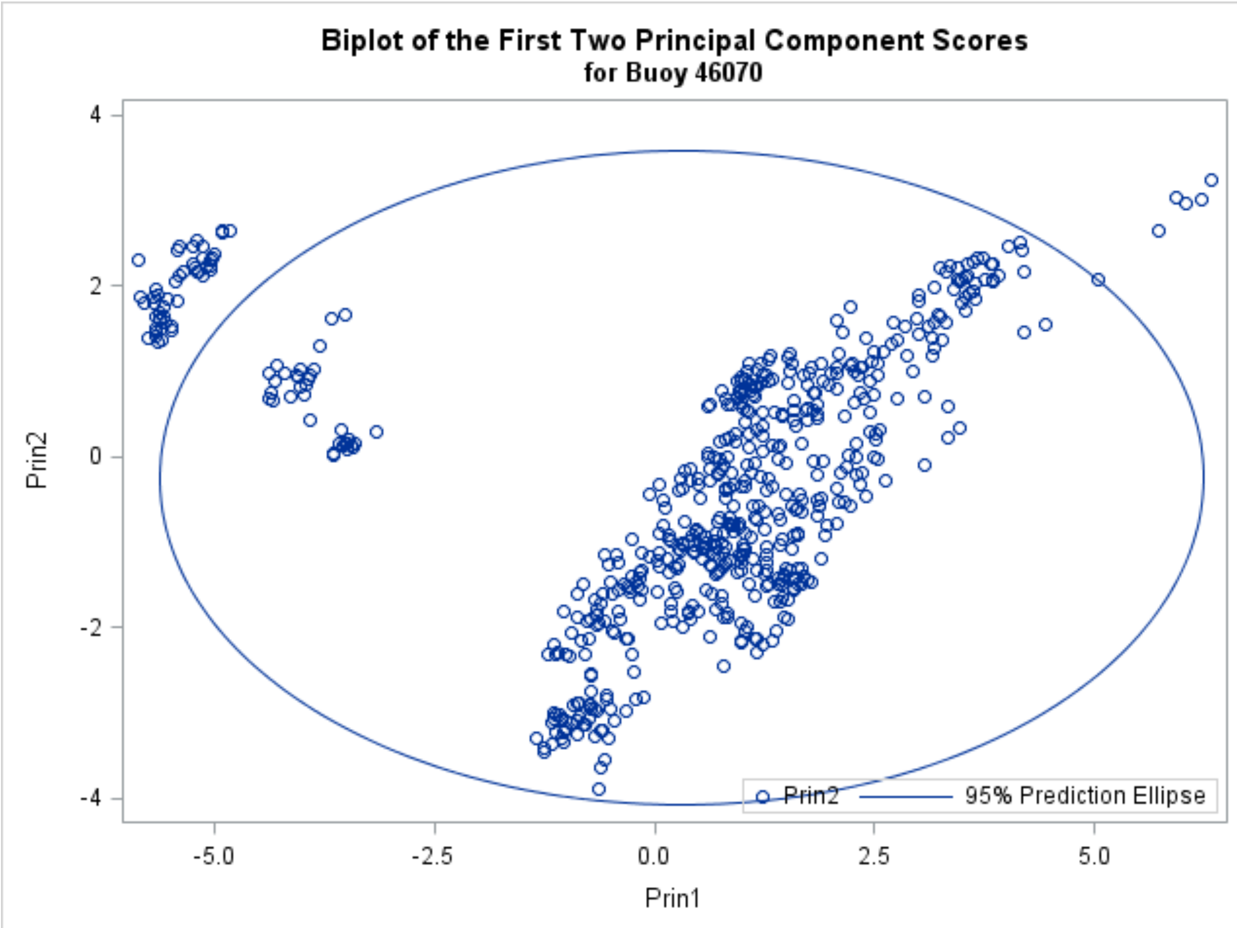

Figure 3: Biplot of the first two Principal Component Scores

Having briefly explored the data graphically, I next ran the statistical tests of interest. Box's M test was run to test the null hypothesis that the data from all of the buoys had an equal variance-covariance structure. With a $\chi^2(165) = 3502.83$, $p < 0.001$, the null hypothesis is rejected. These buoys do not give rise to data with the same variance structure. In other words, data from some buoys is significantly more variable than from others.

A manova test of whether or not the mean vectors were equal was run. Additionally, contrast tests of whether any two buoys had equivalent mean vectors were also run. The results are presented below, in Table 3.

Table 3: Statistical Tests of Mean Vector Similarity for Buoys

Null Hypothesis	Wilk's Lambda	F value	Num DF	Denom DF	p-value
Mean vectors are all equal	0.825	21.30	12	3392.1	<.0001
First and second are equal	0.976	8.04	4	1282	<.0001
First and third are equal	0.949	17.33	4	1282	<.0001
First and fourth are equal	0.906	33.07	4	1282	<.0001
Second and third are equal	0.939	20.72	4	1282	<.0001

Second and fourth are equal	0.886	41.07	4	1282	<.0001
Third and Fourth are equal	0.972	9.29	4	1282	<.0001

It is concluded that none of the mean buoys have the same mean vector. Since neither the mean vectors nor the variance-covariance matrices are equal it must be concluded that the buoys were not sampling the same population in January of 2008.

Focus on a smaller time frame

Given that the weather is highly variable across a month, consider just one day. Below is quick look at January 1, 2008. The database only contains data from two buoys on that day.

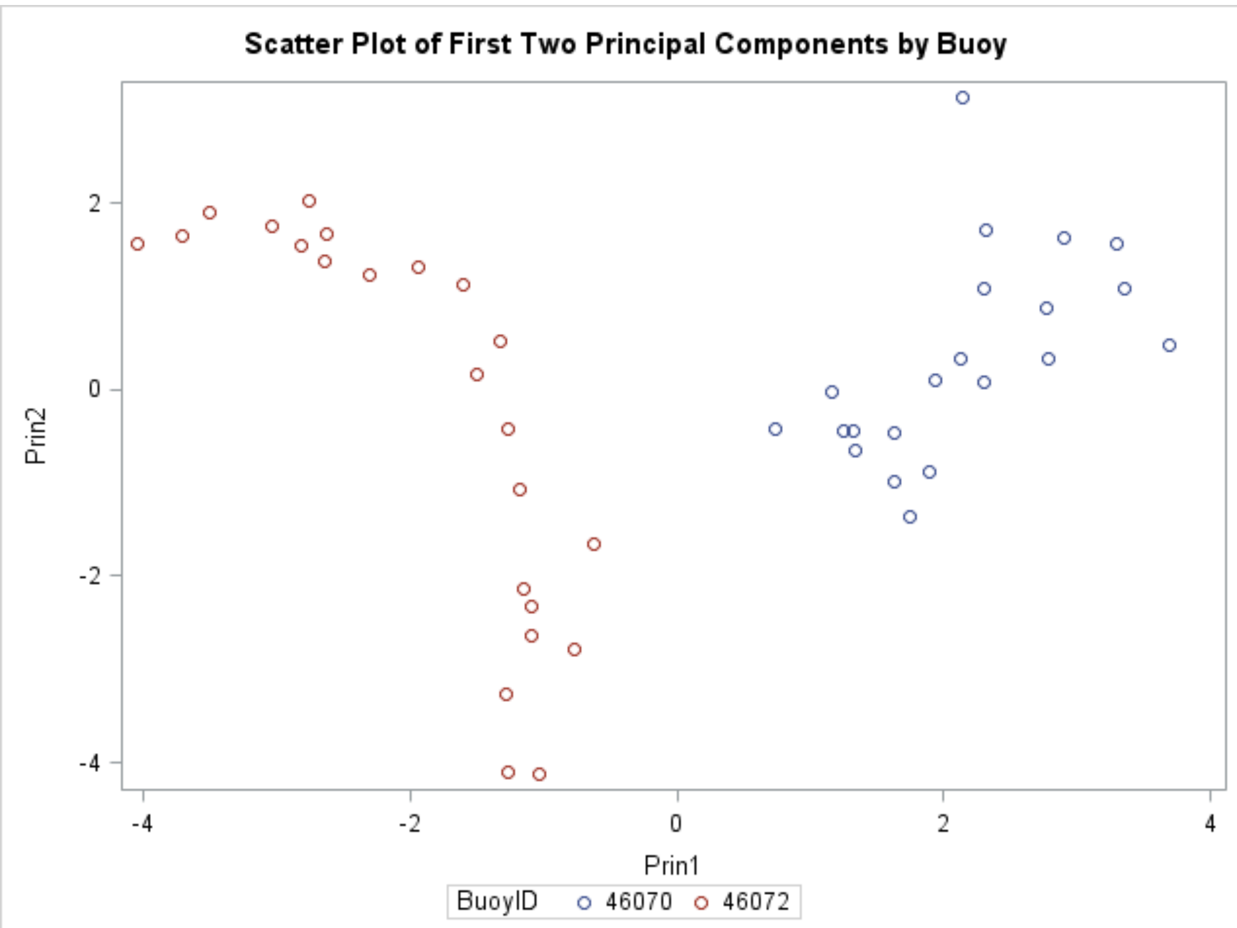

Figure 4: Scatter Plot of the first two Principal Components, by buoy, January 1, 2008.

In Figure 4, the two buoys continue to look distinctly different.

Box's M test indicates that the data from the two buoys does not have the same variance-covariance structure, $\chi^2(55) = 174.40$, $p < 0.001$. Furthermore, a manova test of the equivalence of the mean vectors indicates that the two buoys are significantly different, $F(4,39)=25.27$, $p < 0.001$.

The conclusion is that on January 1, 2008 the weather conditions were not the same at buoys 46070 and 46072.

July 2008

The process was repeated for July 1, 2008, another day for which there was data from the same two buoys (Figure 5).

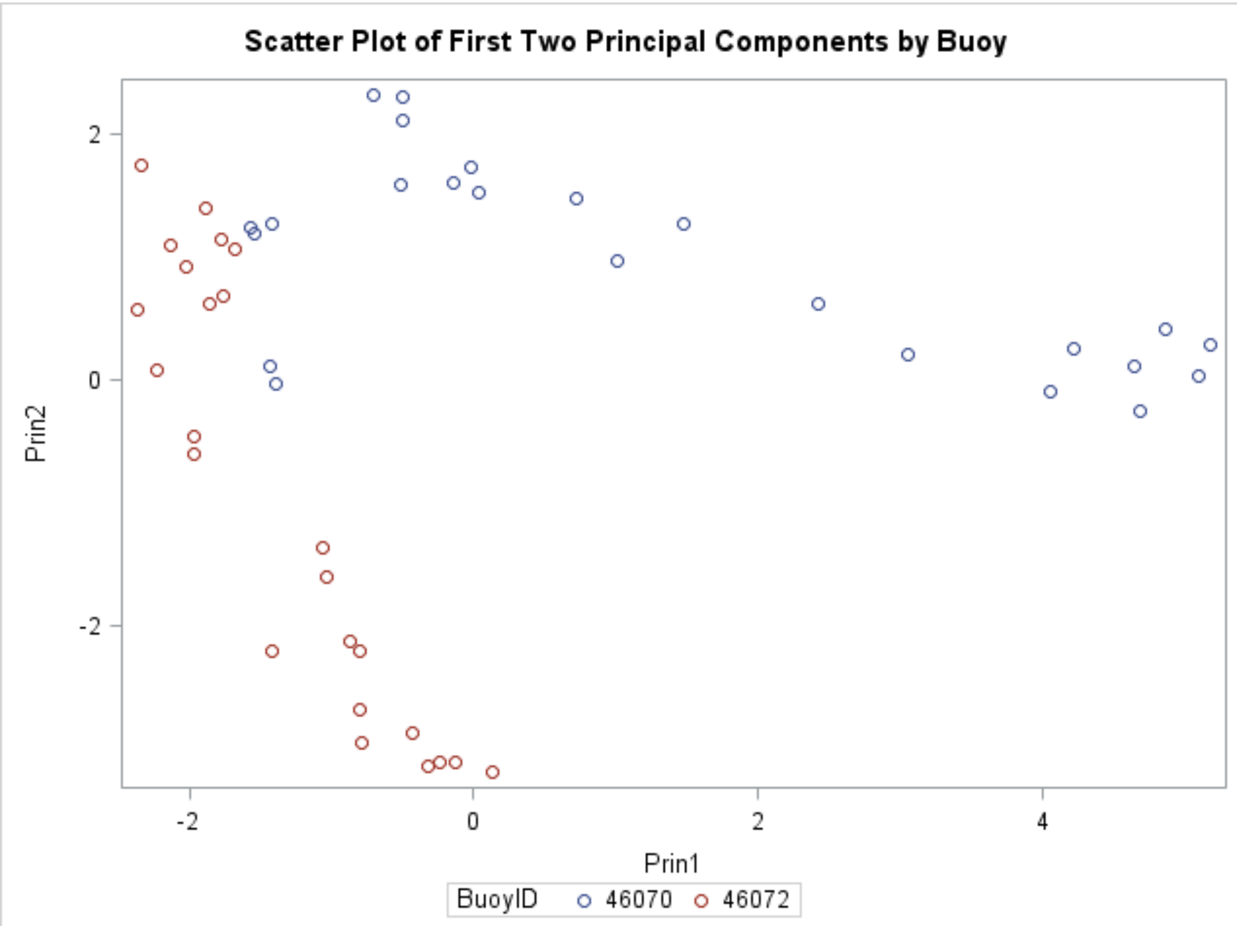

Figure 5: Scatter Plot of the first two Principal Components, by buoy, July 1, 2008.

Box's M test indicates that the data from the two buoys does not have the same variance-covariance structure, $\chi^2(55) = 270.28$, $p < 0.001$. Furthermore, a manova test of the equivalence of the mean vectors indicates that the two buoys are significantly different, $F(4,43)=24.81$, $p < 0.001$. These two buoys do not appear to have the same weather measurements on July 1, 2008.

Consider January 1, 2011, a date for which there is data from the other two buoys, 46073 and 46075.Figure 6 shows scatter plots for the first two principal components of the dimensionally reduced data. It suggests two very distinct populations.

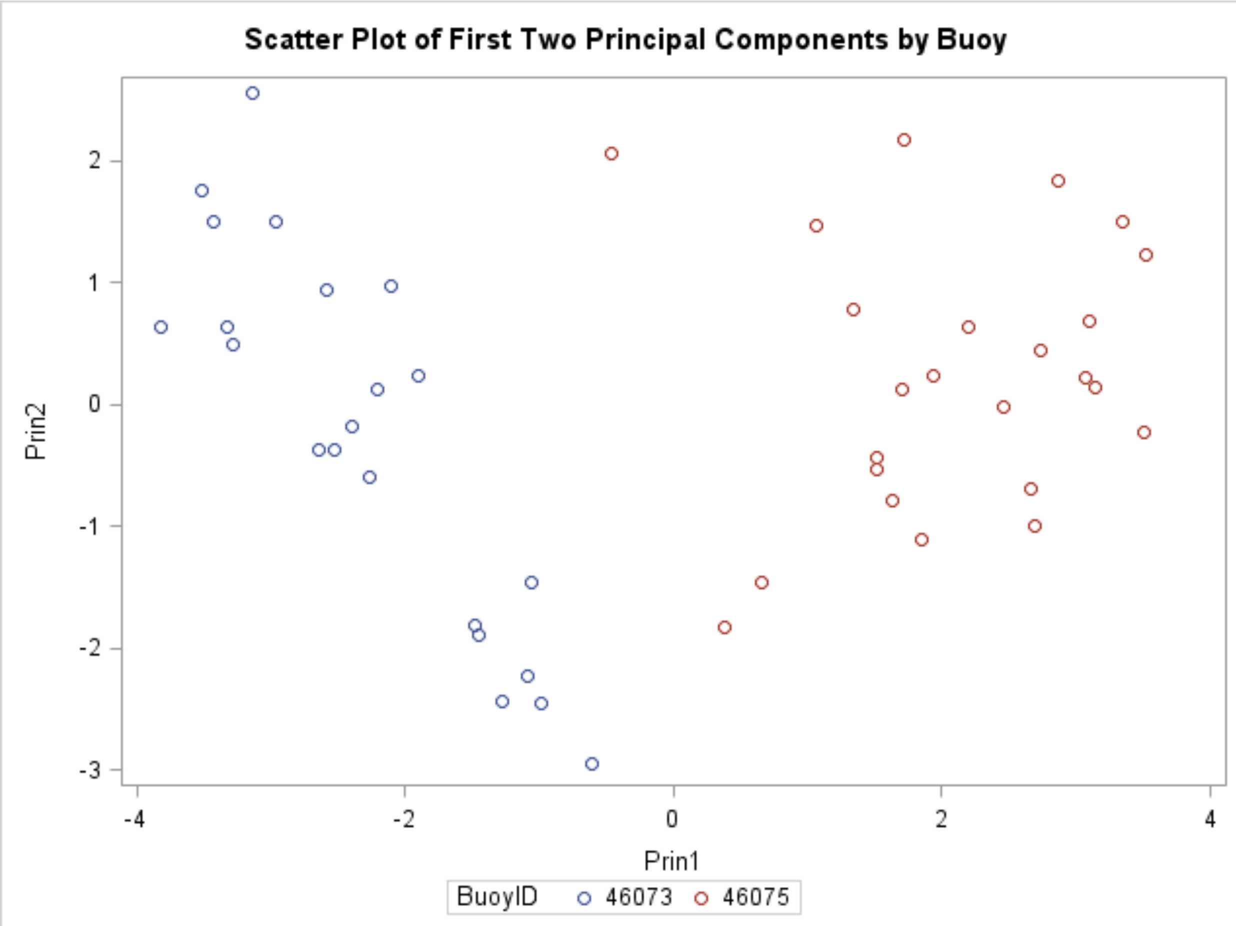

Figure 6: Scatter Plot of the first two Principal Components, by buoy, January 1, 2011.

Box's M test confirms that the variance-covariance structure is not the same for the two buoys, $\chi^2(55) = 188.34$, $p < 0.001$, and the manova test indicates that the mean vectors for the two buoys are significantly different, $F(4, 41) = 31.61$, $p < 0.001$.

The data appears to be conclusive that different buoys give different readings on the weather even on the same day. Assuming each buoy is reporting accurately, this indicates that inferences drawn about one location based upon data received from another would be unreliable.

Airports

A similar analysis was performed comparing the airports. Table 4, which summarizes the sampling frequency at each airport, reveals that, unlike buoys, there are not large gaps in airport data. Therefore, it is easier to compare all three airports over a similar time period.

Table 4: Sampling Frequency of Airports

Month	Year	Number records Adak	Number records Cold	Number records Dutch

1	2005	694	1055	239
1	2006	485	1470	727
1	2007	610	1399	717
1	2008	610	1329	1920
1	2009	699	1220	1413
1	2010	654	1361	1281
1	2011	548	1389	846
1	2012	485	1331	886
2	2005	596	1074	265
2	2006	552	1266	658
2	2007	605	1197	664
2	2008	667	1460	1963
2	2009	581	1238	1263
2	2010	593	1243	1140
2	2011	653	1186	1047
2	2012	603	1380	943
3	2005	714	1236	429
3	2006	722	1234	721
3	2007	601	1441	688
3	2008	716	1919	907
3	2009	458	1973	1265
3	2010	717	1391	1369
3	2011	731	1288	1229
3	2012	648	1505	1182
4	2005	669	1010	623
4	2006	699	1355	718
4	2007	697	1324	712
4	2008	712	1235	1784
4	2009	692	1306	488
4	2010	691	1315	1303
4	2011	717	1354	1289
4	2012	623	1342	1001
5	2005	585	1269	660
5	2006	723	1161	720
5	2007	717	1229	738
5	2008	688	1293	950
5	2009	729	1238	1681
5	2010	741	1332	1194
5	2011	742	1266	1082
5	2012	648	1166	930
6	2005	607	1143	641

6	2006	689	1218	707
6	2007	710	1332	717
6	2008	679	1444	2006
6	2009	686	1152	1284
6	2010	700	1327	1016
6	2011	720	1409	1130
6	2012	626	1137	993
7	2005	665	1201	618
7	2006	716	1223	718
7	2007	722	1353	737
7	2008	721	1320	2147
7	2009	709	1415	1090
7	2010	713	1310	1053
7	2011	743	1491	1127
7	2012	646	1395	961
8	2005	576	1285	651
8	2006	715	1412	733
8	2007	723	1316	734
8	2008	704	1254	2105
8	2009	717	1308	1105
8	2010	719	1292	1091
8	2011	732	1457	1195
8	2012	499	1212	341
9	2005	676	1263	633
9	2006	695	1204	711
9	2007	715	1254	709
9	2008	597	1337	2006
9	2009	693	1227	872
9	2010	692	1100	1020
9	2011	719	1555	1128
9	2012	191	401	329
10	2005	703	1146	652
10	2006	719	1279	726
10	2007	729	1343	736
10	2008	621	1365	1905
10	2009	612	1278	1107
10	2010	715	1321	1210
10	2011	744	1214	1085
10	2012	194	528	356
11	2005	665	1490	635
11	2006	692	1400	612

11	2007	693	1606	719
11	2008	685	1194	2003
11	2009	676	1431	1217
11	2010	696	1363	1103
11	2011	640	1506	1165
11	2012	193	501	341
12	2005	688	1235	644
12	2006	555	1517	739
12	2007	740	1384	744
12	2008	727	1455	1945
12	2009	510	1237	1206
12	2010	703	1581	1513
12	2011	685	1680	1476
12	2012	192	408	281

Because there aren't significant gaps in which data from one or more airports is not available, we can look at summary statistics for the entire data set as a first look for similarities and differences. This summary is presented in Table 5 below. A cursory look at the means suggests the airports are different.

Table 5: Summary Statistics for the Airport Variables							
Airport	N Obs	Variable	N	Mean	Std Dev	Min	Max
Adak	192150	Ceiling_feet	89978	3993.46	3405.23	100	12000
		Visibility_sm	191658	8.3	2.76	0	25
		DryBulbCelsius	189544	4.64	4.24	-13	21
		DewPointCelsius	189425	1.44	4.96	-19	18
		WindDirection	189782	191.08	112.32	0	360
		WindSpeed_Kt	189779	12.19	7.62	0	57.31
		StationPressure_hPa	64595	1007.92	13.76	951.23	1050.12
		Day_illumination_hours	192150	12.45	3.04	7.76	16.71
		Sky_illumination_hours	192150	1.29	0.15	1.12	1.62
		Total_illumination_hours	192150	13.74	3.13	9.12	18.33
Cold Bay	125094	Ceiling_feet	124570	3496.79	3623.23	100	30000
		Visibility_sm	125082	7.3	3.52	0	30
		DryBulbCelsius	125078	2.59	6.18	-19.4	20.6
		DewPointCelsius	125050	0.63	6.53	-22.2	15
		WindDirection	124578	220.19	103.45	0	360
		WindSpeed_Kt	125014	14.3	7.91	0	59.05
		StationPressure_hPa	124770	1002.35	13.48	948.86	1045.37
		Day_illumination_hours	125094	12.26	3.48	7.12	17.42
		Sky_illumination_hours	125094	1.44	0.22	1.21	1.97

		Total_illumination_hours	125094	13.7	3.6	8.65	19.39
Dutch Harbor	103779	Ceiling_feet	102571	3701.59	3152.32	100	25000
		Visibility_sm	103635	8.36	2.79	0	10
		DryBulbCelsius	99867	4.41	4.82	-11	26
		DewPointCelsius	99120	0.75	5.26	-15.6	18
		WindDirection	101832	188.41	115.42	0	360
		WindSpeed_Kt	103198	10.08	7.09	0	52.97
		StationPressure_hPa	101881	1007.2	13.48	945.14	1049.44
		Day_illumination_hours	103779	12.32	3.32	7.39	17.12
		Sky_illumination_hours	103779	1.38	0.19	1.17	1.81
		Total_illumination_hours	103779	13.7	3.42	8.85	18.93

As with the buoy data, a principal components analysis was performed to reduce the dimensionality of the data, so as to capture the information in fewer variables. Graphical exploration of the principal components provides a quick look at whether or not the different populations cluster or not. Figure 7 shows a plot of the first two principal components for all three airports for the date 1/1/2011.

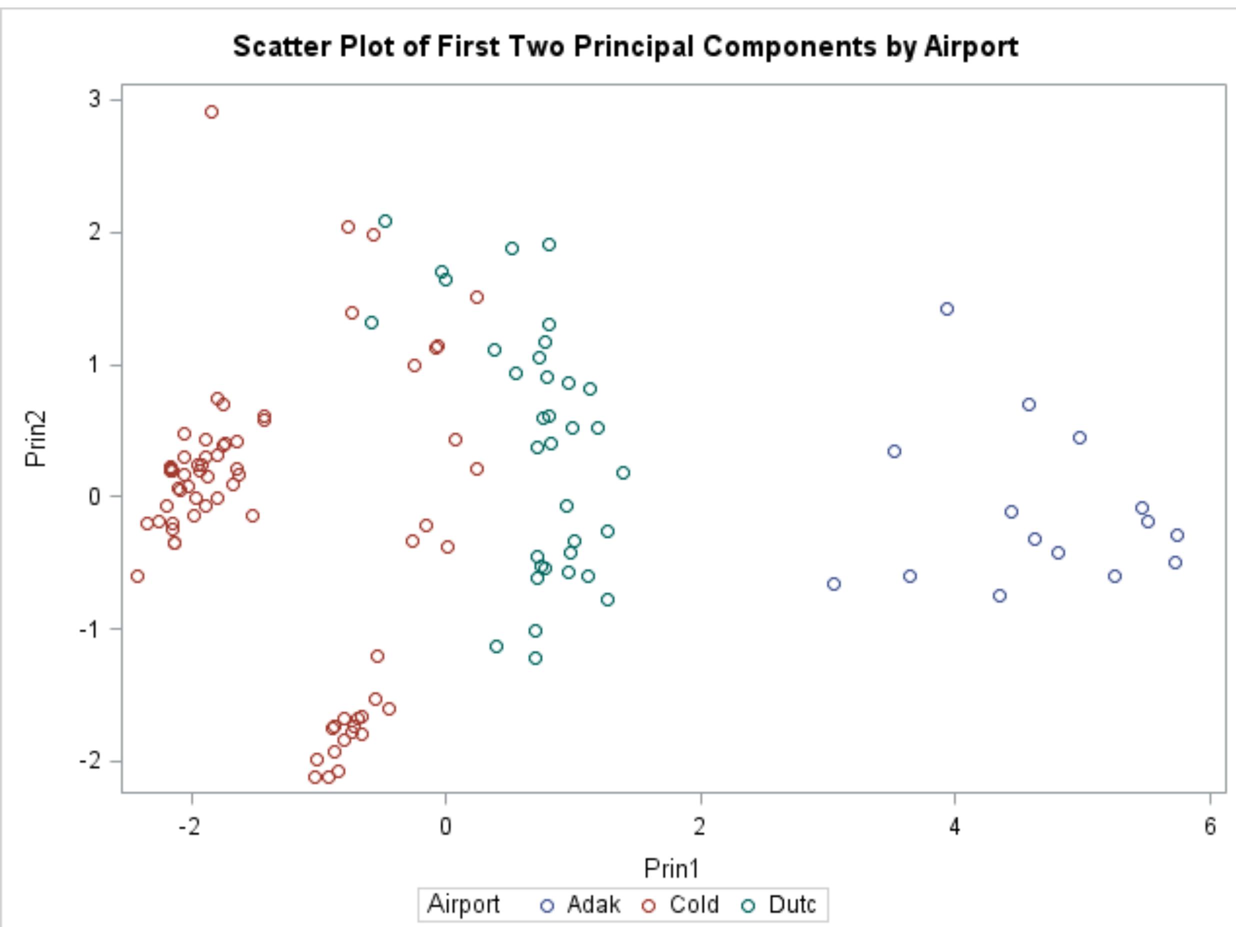

Figure 7: Scatter Plot of the first two Principal Components, by airport, January 1, 2011.

There is clear evidence of clustering within this data, a strong indication that the three airport populations are not equivalent.

The original variable vectors were tested for equivalence using the statistics discussed above. The test of the null hypothesis that the variance-covariance matrices for the three airports is the same must be rejected, $\chi^2(56) = 516.55$, $p < 0.001$, and the manova test of equal mean vectors for the airports also finds significant differences (see Table 6, below).

Table 6: Statistical Tests for Similarity of Mean Vectors for Airports, January 1, 2011

Null Hypothesis	Wilk's Lambda	F value	Num DF	Denom DF	p-value
Mean vectors are all equal	0.021	94.26	14	222	<.0001
First and second are equal	0.060	248.96	7	111	<.0001
First and third are equal	0.132	104.17	7	111	<.0001
Second and third are equal	0.162	82.16	7	111	<.0001

It must be concluded that on January 1, 2011 the weather readings for the three airports did not come from equal populations.

For good measure another date, July 1, 2009, was also studied. Once again a graphical scatterplot presentation of the first two principal component variables is shown below. It indicates distinct differences in the data from the three airports.

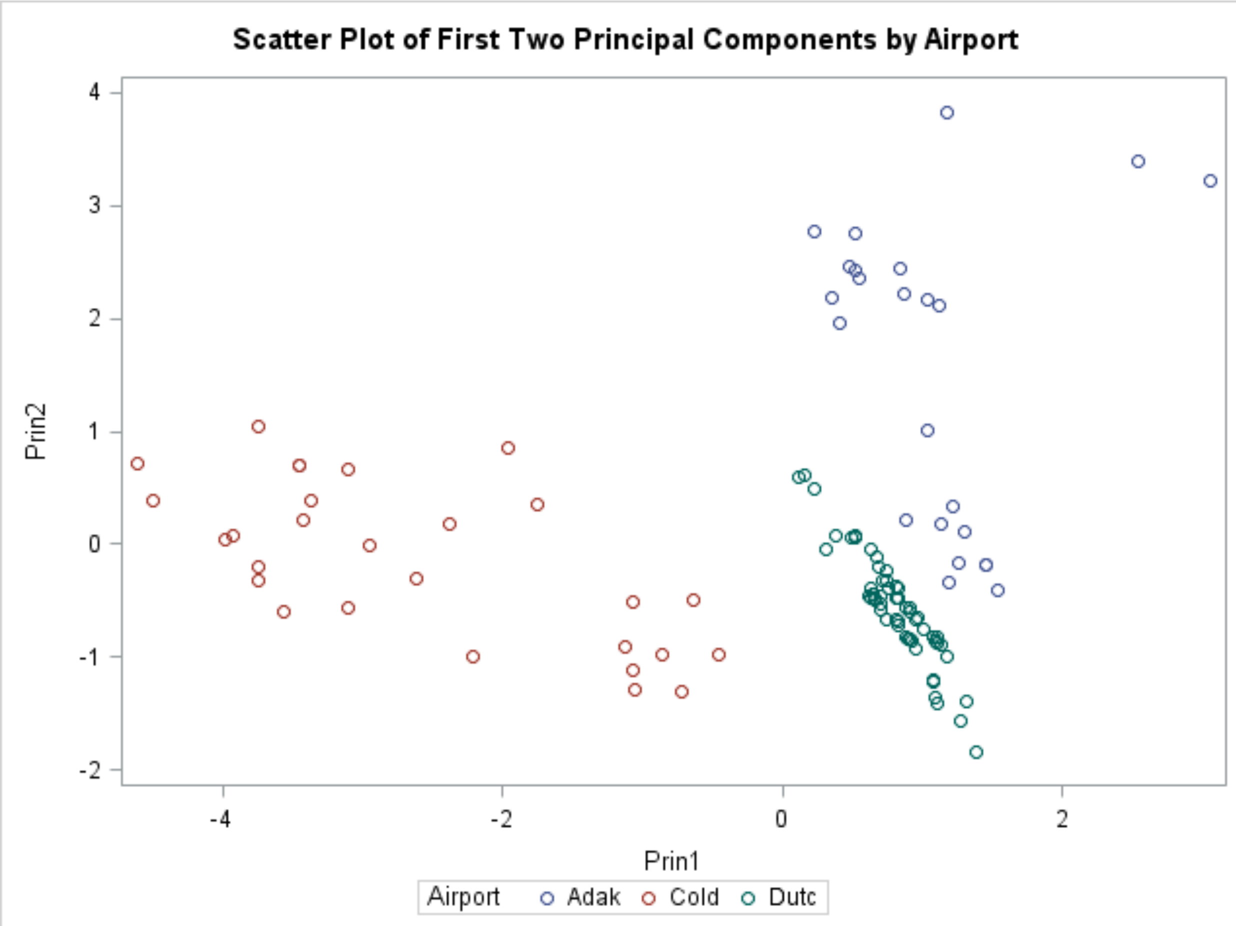

Figure 8: Scatter Plot of the first two Principal Components, by airport, July 1, 2009.

The statistical tests for equivalence were repeated. A test of homogeneity of covariance matrices rejected the null hypothesis that they were equal, $\chi^2(56) = 2594.75$, $p < 0.001$. The manova test results shown in Table 7 confirm that the mean vectors are not equal for any combination of the three airports on the July date studied.

Table 7: Statistical Tests for Similarity of Mean Vectors for Airports, July 1, 2009

Null Hypothesis	Wilk's Lambda	F value	Num DF	Denom DF	p-value
Mean vectors are all equal	0.004	205.66	14	200	<.0001
First and second are equal	0.044	310.42	7	100	<.0001
First and third are equal	0.137	89.76	7	100	<.0001
Second and third are equal	0.033	417.77	7	100	<.0001

There is no evidence that the reported weather at the three airports is from the same population, and inferences about one airport based upon data from another are not warranted.

Conclusions:

This work was motivated by the suggestion that missing information from various weather collection points could be filled in using information from surrounding data collection points. Inherent in this proposal is the assumption that the collection points are actually sampling the same weather populations. If that were so, the vectors containing the mean values for the variables of interest would be expected to be statistically equivalent as would their variance-covariance structures. To test this, subsets of the entire data set representing short time periods for which data from all sources was available were tested for equivalence. Manova regression was used to test for equivalence of the mean vectors, and Box's M test was used to test for equivalent variance-covariance structures.

The tests were repeated for multiple summer and winter periods. In all periods studied, it was concluded that the mean vectors for the four buoy populations and the mean vectors for the three airport populations were significantly different. Furthermore, the variance-covariance matrices for the various data collection sites were also found to be significantly different.

Scatter plots created after reducing the dimensionality of the problem through the use of principal components analysis graphically confirmed the observation that the data populations were statistically unique. Based on these findings, imputation for missing values pertaining to one data collection point using data