# Analyzing NHL Goalie Stats (03-04—07-08) Using the Self-Organizing Map

**Chuck Crittenden**
**Data Mining**
**12/17/2008**

"In hockey, goaltending is 75 percent of the game.
Unless it's bad goaltending. Then it's 100 percent of the game,
because you're going to lose."
~ Gene Ubriaco (NHL forward)

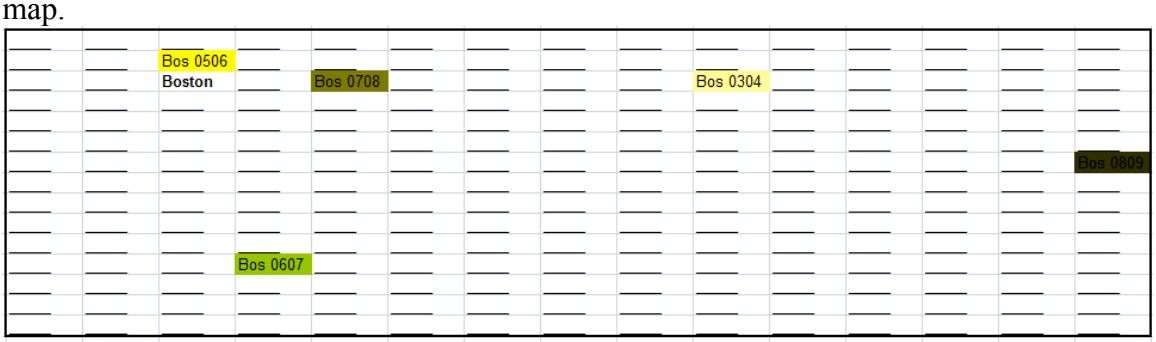
**Executive Summary**

It's always been said that great goaltending will lead you to the playoffs. Last year I proved it. (Crittenden – NHL Goalie SOM, Aleshunas). I'd like to use the knowledge I gained last year to use the Self-Organizing Map to monitor the progress of each NHL team over the course of four seasons. Last year I used the Self-Organizing Map to attempt to find natural clusters of different levels (based on average standings) of NHL teams together using their average Goals Against Average (GAA), Save Percentage (SV%), Goals Allowed (GA), Goals Scored (GF), and Goal Differential (DIFF) over the seasons 2003-2004 through 2007-2008 (omitting the 2004-2005 season since their was a strike for the duration of that season). This year I will further my work using the same data.

The team name will become my label with which to map. In short the Self-Organizing Map uses a map with a pre-defined size randomly filled with attributes. The instances of data are compared with each individual point on the map using the Euclidean distance. Whichever point on the map the instance is nearest to it is made into that map point. At this point the program also trains the instances around it using competitive learning. This process repeats for every instance. The algorithm repeats for a predetermined number of repetitions. I want to have my map as clustered as possible without having too large of a map. If I had too large of a map, the results would be so spread out that the results wouldn't be as clear.

Last semester using NHL goaltending data (retrieved from NHL.com and Yahoo Sports) with the Self-Organizing Map, I was able to come up with a map that had a good spread for each of the levels in the standings. The high level teams mapped together, with the next level following below it, and so on until the low level teams mapped into the opposite corner of the map.

Now I want to take this a step further and see how teams progressed throughout the four seasons. The team's map that I think best expresses what I wanted to show is Boston's map.



Their point totals starting in 03-04 and going to 07-08 were 104, 74, 76, and 94. Notice how Boston has 104 and is near the top right. Then as their point levels drop to the 70s they move further left. But in 07-08, they are significantly better in the standings with 94 points and it shows in the big jump up the map. Also, so far this season Boston is number two in the league. They are once again near the top right of the map.

Also, I wanted to test the SOM a little bit to find out which technique of mapping separate teams would yield the best results. I found that using a seed for randomization caused the maps to be a little sporadic and not necessarily provide consistent results. Creating a single map using all of the data allowed me to map different teams or different years quickly and easily. Creating a single map was much superior to using a seed.

The findings are that the Self-Organizing Map is a very good way to monitor the progression of an NHL team based on GAA, SV%, GA, GF, and DIFF. The one major thing that must be remembered when using the Self-Organizing Map is that if the program is re-run, the results will be different, because of the random initialization unless a seed is used. Adjusting the number of repetitions or the dimensions of the map will also result in a different map. The Self-Organizing Map did a very good job of chronicling the progress of each team and providing more insight into how much of an effect goaltending has on a team's standings.

**Problem Description**

I intend to further my efforts on a problem I worked on previously. The problem that I will attempt involves using the Self-Organizing Map (SOM) to try to effectively cluster National Hockey League (NHL) goaltending statistics from the 2003-2004 season through the 2007-2008 season. My previous work classified the data into five different levels of hockey teams. These five levels were decided based on each team's average standings during these particular seasons. This does not include the year of 2004-2005, because of the NHL lockout during that particular season. The resulting map from my last project was this:

| | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Pittsburgh | x | x | x | Carolina | x | x | x | x | LosAngeles | x | x | StLouis | x | Chicago |
| x | x | Toronto | x | x | x | Florida | x | x | x | x | x | x | x | x |
| x | x | x | x | x | x | x | x | NewYorkI | x | Washington | x | x | x | x |
| Vancouver | x | x | TampaBay | x | x | x | x | x | x | x | x | x | Phoenix | x |
| x | x | x | x | x | x | x | x | x | Atlanta | x | x | x | x | x |
| x | x | x | x | x | x | x | x | x | x | x | x | x | x | x |
| NewJersey | x | x | Calgary | x | x | x | x | x | x | x | x | x | Boston | x |
| x | x | x | x | x | x | x | x | x | x | x | x | x | x | x |
| x | Colorado | x | x | x | x | x | Montreal | x | x | Edmonton | x | x | x | x |
| x | x | x | x | x | x | x | x | x | x | x | x | x | x | x |
| Buffalo | x | x | x | x | x | x | x | x | x | x | x | x | x | Philadelphia |
| x | x | x | x | x | x | Anaheim | x | x | x | x | x | x | x | x |
| x | Ottawa | x | x | x | x | x | x | x | Minnesota | x | x | x | x | x |
| x | x | x | SanJose | x | Dallas | x | x | x | x | x | x | x | x | x |
| Detroit | x | x | x | x | x | x | Nashville | x | x | x | x | x | NewYorkR | x |

The different levels of teams almost layer on top of each other showing that the groups did in fact cluster together.

Currently I would like to use the SOM and the same data set to analyze how a team progressed through the four seasons. I also would like to see if I can make some sort of future prediction based on the past seasons and the current year's data so far. One other problem that I would like to address is how I will ensure all of the maps will have the same base. (Meaning the best teams are always in the same place of the map.)

**Analysis Technique**

The algorithm I am going to use to cluster the teams is the Self-Organizing Map or SOM.

The Self-Organizing Map (a type of artificial neural network) is a method of finding clusters in high dimensional data and showing them in a 2-dimensional map. The program first randomizes instances into the points on the map (randinit). The points are randomized within the range of each attribute. Next the program takes a specific instance (p) and tests it against each of the points on the map (q) and finds the one point where the Euclidean distance between the two is smallest.

$$\sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \cdots + (p_n - q_n)^2} = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2}.$$

(Euclidean Distance, Wikipedia) The point p is placed into that point (q) and q is trained to more closely resemble the point (p) it was being compared to. The algorithm then trains the surrounding nodes with less training the further from the original node it is. This process repeats for all the instances for 'rlen' number of times. This process along with what else is needed is described in detail in the following paragraphs. I will walk through setting up the map that uses all 4 seasons. (Stepping through the algorithm, Wikipedia)

The first thing that is necessary to do is to organize the data within the spreadsheet containing the data we retrieve. We know that we want to use the teams as our labels. The format required for the executable programs is that the labels must be at the end of numerical data. In order for our map to contain the full name of each team, we must remove all spaces from the team name.

The next step is deciding what attributes are going to be used to find any clusters. I collected data for NHL goaltenders from the seasons 2003-2004 through 2007-2008 with 2004-2005 being omitted due to the strike. The data I retrieved were the average team Goals Against Average (GAA), the average team Save Percentage (SV%), the team's goals allowed (GA), the team's goals scored (GF), and the team's goal differential (DIFF) for each year. Finally I took the average GAA, SV%, GF, GA, and DIFF for each team for the four seasons. This data is what I used in SOM. (Hockey Data, NHL & Yahoo Sports)

I chose to use GAA and SV% for my data, because in hockey these are the two most important statistics for goaltenders. GAA measures on average how many goals a goalie allows in a game. It is calculated by:

$$\frac{\text{Goals Allowed}}{\text{Number of Minutes Played}(1/60)}$$

SV% measures how many saves a goalie will make out of 100 shots. It is calculated by:

$$\frac{\text{Goals Allowed}}{\text{Shots Allowed}}$$

GF is the number of goals scored, and GA is the number of goals allowed.
The last statistic I chose to use is DIFF. DIFF is calculated by:

Goals Scored – Goals Allowed = DIFF

The reason I am including GF, GA, and DIFF is to attempt to remove any noise from the data. Teams that have a great goalie but a bad offense and do not make the playoffs would map higher on the map if these were not included. It is the same principle for teams with a very high scoring offense and a mediocre goaltender. I hypothesize that this will even my map out to cluster teams correctly.

The next step is converting it to the proper format for the executables. For each .dat file (essentially a .txt file saved as .dat), the first line should be the number of attributes (not including the label), 5 in this case. After that comes each team, with GAA, SV%, GA, GF, DIFF, and team name. And so on until the end with no carriage return after the last instance. This is the labeled .dat file (say nhl_label.dat). We also want an unlabeled .dat file (say nhl.dat), which is the same file except for all of the labels are not in the file.

After this, we want to set up the proper .bat file (say nhl.bat) to execute the programs properly. The three executables referenced are randinit, vsom, and vcal. (Self-Organizing Map (SOM), Aleshunas). The .bat file should look something like this:

```
randinit -din nhl.dat -cout nhl.cod -xdim 15 -ydim 15 -topol rect -neigh bubble -rand 0
vsom -din nhl.dat -cin nhl.cod -cout nhl.cod -rlen 10000 -alpha 0.05 -radius 15
vsom -din nhl.dat -cin nhl.cod -cout nhl.cod -rlen 1000000 -alpha 0.02 -radius 5
vcal -din nhl_label.dat -cin nhl.cod -cout nhl_label.cod
```

In this example, ‘nhl.cod’ is a codebook passed between the programs, ‘xdim’ and ‘ydim’ are the dimensions of the map. This shows I am going to be using a 15x15 map. This size map allows room for a large amount of the teams to be mapped and it is not so large that it is hard to analyze. ‘rlen’ tells the program how many times to run that algorithm. ‘radius’ refers to the training radius for each instance. ‘nhl_label.cod’ is the output codebook with labels. To run the program, simply start this .bat file after making sure it is in the same folder as the ‘.exe’s and the proper .dat.

To map these results with the labels, you can either do it by hand or use som_mapper.exe. However with the executable you need to make sure there is not a carriage return at the end of the file. If there is, it is not a difficult fix. You simply need to perform some maintenance on your map after the data has been mapped. (Namely changing the BLANK in the bottom right corner to _____.) You also need to create a file named ‘control.dat’ in order to run som_mapper.exe. Its contents should be:

0 nhl_label.cod nhl_output.txt

If you wanted to map the resulting attribute, you can change the 0 to a 1, 2, 3, 4, or 5 in this example depending on which attribute you wanted. The 1 corresponds to GAA, the 2

to SV%, the 3 to GA, the 4 to GF, and the 5 to DIFF. nhl_output.txt placed into an excel file as space delimited would give you a map that you could then color-code as you desire.

When taking this a step further and knowing I want to create and analyze multiple maps from the same data, I need to figure out how I can get them to map in the same general direction. In this example the best teams will always map to the top right corner, because that is how they were randomly assigned.

Since the map is ready to receive labels after the second vsom executable has been run, we can run vcal with whatever labeled data set we would like. I will choose to run it once for each team to watch how they progressed through the years. I will also run it once for each season to see how the maps as a whole progress.

There is another way we could perform this same task. We could run the entire program using a random seed. This in theory would ensure that all of the maps would begin in the same position. Soon after doing this for each season's data, it was becoming very time consuming, and the map was not as clearly defined as the other process. It was time consuming, because you had to run al of the executables not just one. And for every new map you had to change the .bat files to represent the change.

The last idea I wanted to test out is whether or not we could possibly predict who would be the playoff contenders based on the current season's data so far. To attempt to accomplish this, I mapped the data for the 08-09 season (through 12-9-08) onto the data for the past four seasons assuming the past four seasons were representative of how this season would finish.

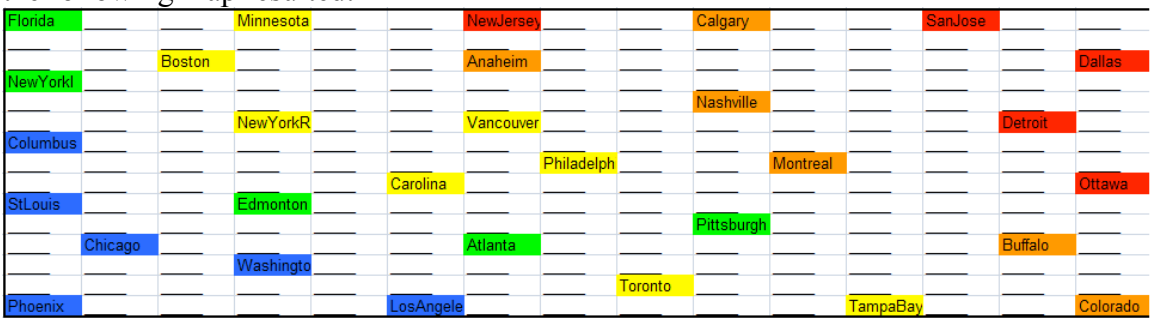
The only issue I can see arising is if the data being input into the program is not in the correct format. Also the user of the program must be able to make a .dat file and use an .xls file to analyze the data. If the user isn't able to, this poses an issue as to whether or not they can fully complete the problem stated.

**Assumptions**

Each team's goaltending statistics are accurate representations for all goaltending statistics in NHL history. This is necessary if we are going to make the assertion that these statistics can always decide who would be in the playoffs.
The data I retrieved was accurate information.
The algorithm performed correctly.
For the prediction aspect of my problem to be accurate, I must assume that the previous four years of data will be an accurate representation of where the best teams will lie on the map. I must also assume the first portion of the season will be representative of the rest of the year.

**Results**

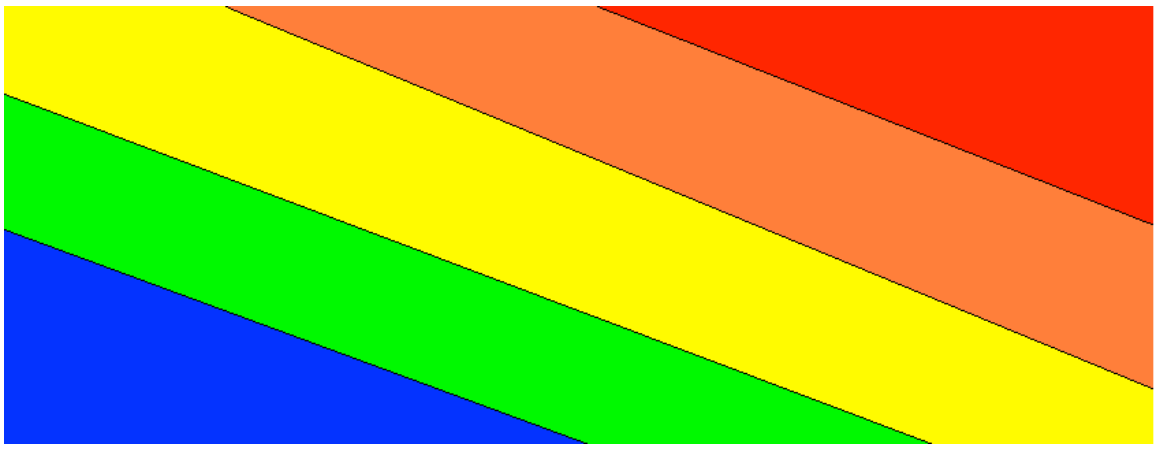
After running the five statistics (for all four seasons) through the Self-Organizing Map, the following map resulted:




There are five different levels of teams in this particular map. They are color-coded as follows: Red are the teams with the highest average standing, Orange, Yellow, Green, and Blue follow respectively, with Blue being the lowest set of teams. (Stats, NHL & Yahoo Sports) The list of each team, their average standing during the four seasons, and their grouping are:

| | overall standings | |
|---|---|---|
| Detroit | 1.25 | high |
| San Jose | 5.75 | high |
| New Jersey | 7 | high |
| Ottawa | 7.25 | high |
| Dallas | 7.75 | high |
| Buffalo | 10.5 | medhigh |
| Anaheim | 10.75 | medhigh |
| Nashville | 10.75 | medhigh |
| Calgary | 11.75 | medhigh |
| Colorado | 12 | medhigh |
| Montreal | 12.5 | medhigh |
| Philadelphia | 14 | med |
| Vancouver | 14 | med |
| New York R | 14.5 | med |
| Minnesota | 14.75 | med |
| Carolina | 15.25 | med |
| Tampa Bay | 15.75 | med |
| Toronto | 16 | med |
| Boston | 16.25 | med |
| Pittsburgh | 18 | medlow |
| Edmonton | 18.75 | medlow |
| Atlanta | 20 | medlow |
| New York I | 20.25 | medlow |
| Florida | 22 | medlow |
| St. Louis | 23.5 | low |
| Washington | 23.5 | low |
| Los Angeles | 24.5 | low |
| Columbus | 25.5 | low |
| Phoenix | 25.5 | low |
| Chicago | 25.75 | low |

The resulting map turned out to be about what was expected. All five of these levels for the most part mapped together like this:

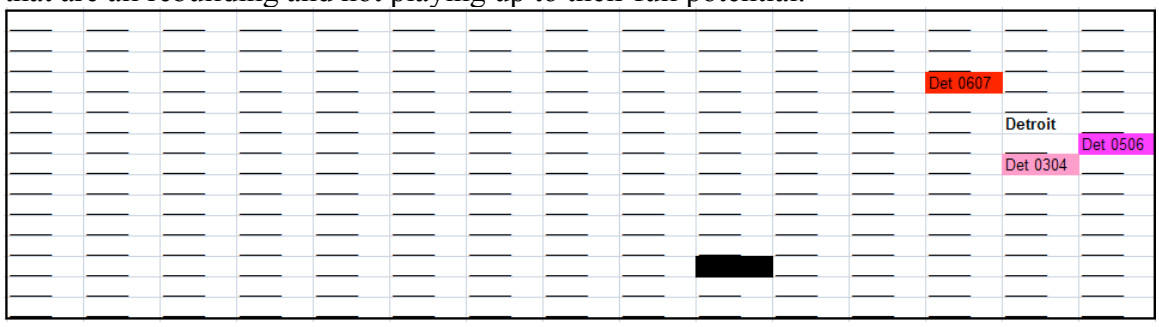


All five of the highest level mapped to the top right corner: Detroit, San Jose, New Jersey, Ottawa, and Dallas. The next highest, Orange, is almost layered underneath the Red level. Yellow seems to be in between Red/Orange and Green/Blue. Green and Blue are separated but are both located in the bottom left of the map.

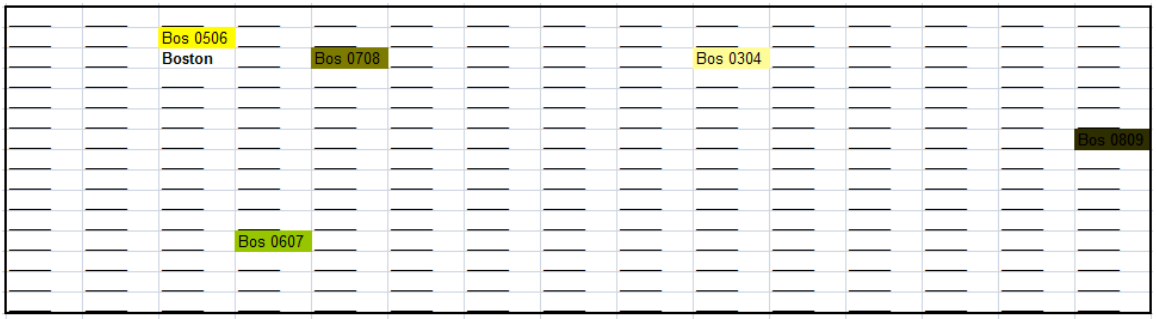
There is some overlapping within the map, but that is discussed in the Issues section, which is forthcoming.

After putting all of the teams through the SOM to see how they progressed, a few interesting paths emerged. The years in the following maps start with the lightest color for the first season and get darker for each successive season, ending with the current season (08-09).

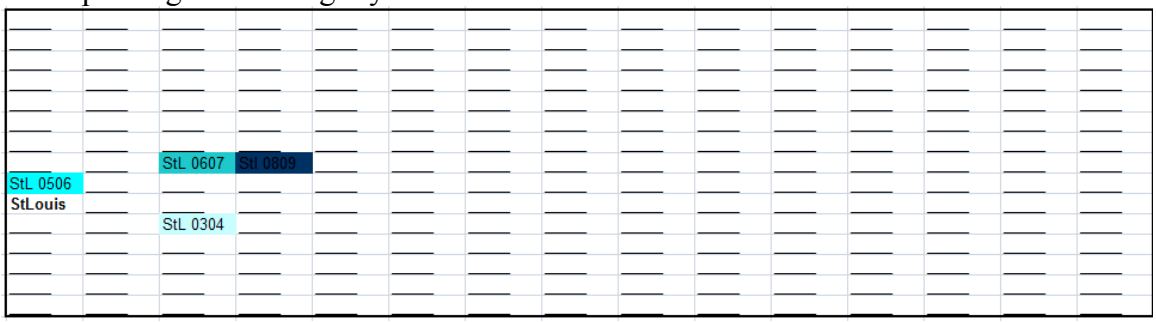
For Detroit, they had a significant goalie in their starting role for the past 4 seasons. However, this season Osgood and Conklin have had to split the role. This may have possibly leaded to the drop in their map position. Though, currently they are second in the conference. This discrepancy might be due to them playing in a division with teams that are all rebuilding and not playing up to their full potential.




In Boston, their standings in the league have fluctuated quite a bit, and the following map shows it. In 03-04 they had 104 points and were fourth in the league, 05-06 they dropped to 26$^{th}$ and 74 points, 06-07 saw them in 23$^{rd}$ with 76 points, 07-08 they were 15$^{th}$ with 94 points, and so far in 08-09 they are in second place in the league. This map shows that they started off near the top right but drifted downward but are getting better with last season and the current one.

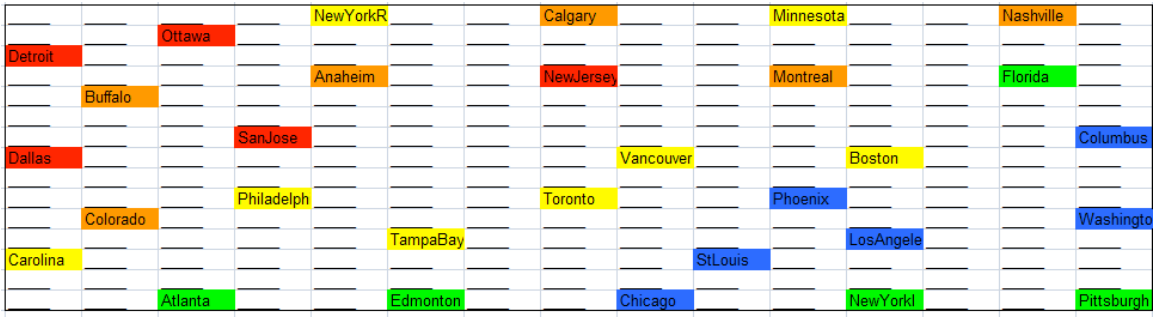



For any St. Louis Blues fan it is common knowledge that the Blues have been staying at about the same place in the standings for the past few years. Their map shows just that. They have not gained nor lost a significant amount of ground. However it appears they are improving ever so slightly.

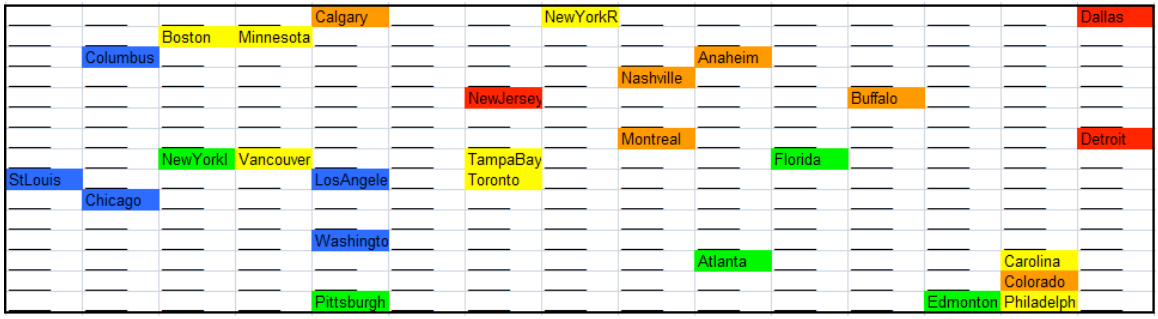



While I attempted to recreate the making of separate maps for each year using a seed for the randomization, I ended with two sets of maps that differed more than I would have hoped. For example in both random maps, I was planning on the top teams mapping together in the same area, but in 05-06 they map to the top left and in 03-04 they map to the top of the map. Also in 06-07 they would map to the left of the map. While none of these maps have a very significant separation, they all do separate the reds and the blues fairly well, which only backs up the point that goaltending statistics do make a difference.

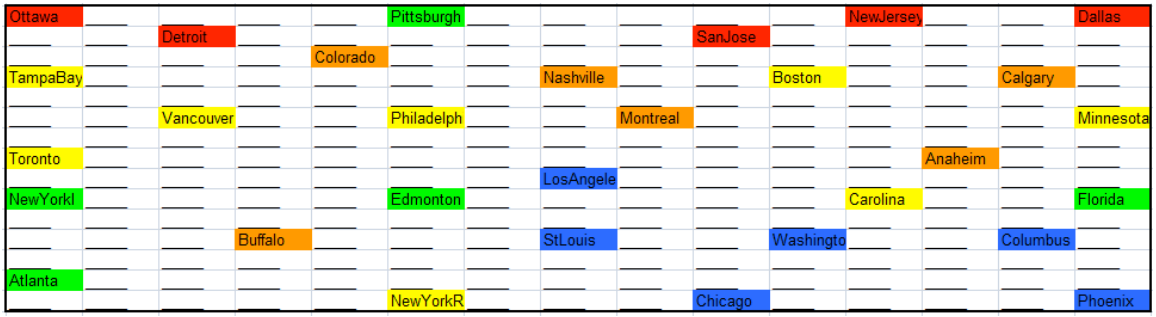
**05-06Random**



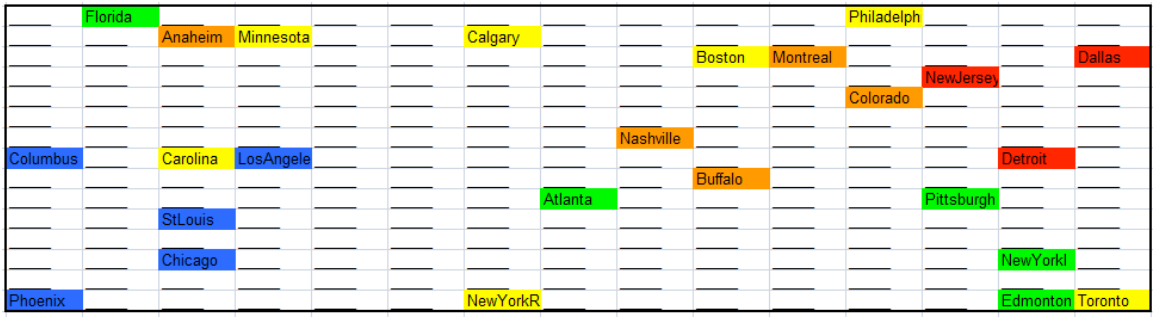
**05-06**




**03-04Random**

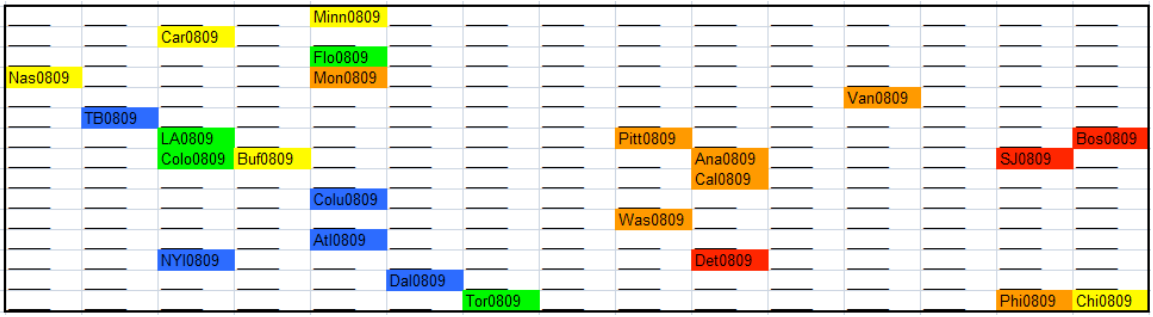



**03-04**




Seeing how the seeded maps differ amongst themselves, I am confident that making one complete map of all the data and attach labels was the best choice for monitoring the progression of the teams.

The last point I wanted to attempt was to see if I could predict how the current season would turn out based on past data. I won't know until the end of the season, but it appears that Vancouver, Boston, San Jose, Pittsburgh, Anaheim, and Calgary will be the teams to beat this post season.




**Issues**

Overall, the only real issues that were encountered happened in the final maps themselves. As you can see in the first map (of all four seasons), there are a few teams that are a bit out of place. The most apparent of these is Pittsburgh, but if you look at the overall standings list, Pittsburgh is the closest Green team to being a Yellow team. The other is Boston, and once again if you look at the overall standings, Boston is the closest Yellow team to being a Green team. These small overlaps are not a problem, since the rest of the map worked out fairly well.

The main issue in the team-specific and year-specific maps is that the data for the entire four seasons may not be completely representative of each single team or year. However, this is an assumption that must be made within the scope of my project.

The issue in seeding the randomization is that the maps didn't turn out as separated as the other ones did. It is also much more time-consuming and not worth the effort to only finish at a clouded result.

The issue in predicting how this season will end is that anything can happen between now and the rest of the season.

**Appendices**

For more information regarding Euclidean Distance, visit:
http://en.wikipedia.org/wiki/Euclidean_distance

For more information regarding SOM, visit:
http://en.wikipedia.org/wiki/Self-organizing_map
or
http://www.cis.hut.fi/teuvo

**References**

Aleshunas, John. Retrieved Apr. 17, 2008. "Self-Organizing Map (SOM)" from:
http://mercury.webster.edu/aleshunas/MATH%203210/MATH%203210%20Source%20Code%20and%20Executables.html

Aleshunas, John. Retrieved Dec. 9, 2008. "Crittenden – NHL Goalie SOM" from: http://mercury.webster.edu/aleshunas/Support%20Materials/SOM/Crittenden%20-%20NHL%20Goalie%20SOM.doc

Goaltender's Annex. Retrieved May 5, 2008. Ubriaco Quote from: http://www.angelfire.com/sk/goalieannex/quotes02.html

NHL.com. Retrieved Apr. 16, 2008. "Goalie Statistics and Team Standings" from: http://www.nhl.com/nhlstats/app

NHL.com Retrieved Dec. 9, 2008. "08-09 Goalie Statistics and Team Standings" from: http://www.nhl.com/nhlstats/app

Yahoo Sports. Retrieved Apr. 16, 2008. "Goalie Statistics and Team Standings" from: http://sports.yahoo.com/nhl/teams/___/stats (Replace ___ with each team's abbreviation).

Yahoo Sports. Retrieved Dec. 9, 2008. "08-09 Goalie Statistics and Team Standings" from: http://sports.yahoo.com/nhl/teams/___/stats (Replace ___ with each team's abbreviation).

Wikipedia. Retrieved Apr. 17 2008. "Stepping through the Algorithm" from: http://en.wikipedia.org/wiki/Self-organizing_map - Stepping_through_the_algorithm

Wikipedia. Retrieved May 6, 2008. "Euclidean Distance" from: http://en.wikipedia.org/wiki/Euclidean_distance