Does Pace Correlate with Success?

Dilpreet Grewal and Justin Mahun

Simon Fraser University

Greg Baker

CMPT 353: Computational Data Science

December 6, 2024

When watching most sports, there is one aspect of play which keeps us hooked: pace. The notion of pace in sports is something that can be thought of in the general sense as a high frequency of important events. Specifically, in the context of this report and many previous explorations and ideas surrounding pace in sports, a high frequency of important events on offence. Watching a high-tempo and high-scoring team makes for better entertainment than a dominant defence, but does playing with a high pace result in better performance?

Previous Exploration of Pace in Sports

Pace in open-play sports such as hockey and soccer is challenging to quantify, due to the fluid nature of these sports, where offence and defence often intertwine. However, a sense of pace in these sports can often be gathered by watching. For example, the philosophy of 'tiki-taka' in soccer which emphasizes possession and short passes above all else, which consequently 'slows' the game down. On the other hand, teams can choose to play 'counter-attack', which trades possession for aggressive runs towards the opponent goal as soon as the ball is gained. Therefore, it seems that location based metrics, such as the location and velocity of the ball and/or players would be ideal to quantify pace. For quantifying pace in soccer, one investigation defined an 'attacking pace' (which was found to correlate with 2 additional shots on goal per game), as the ratio of actions which move the ball closer to the opponent goal / total possession time (Dona & Swartz, 2023). This conception of pace was similarity applied by a primary researcher of this paper, towards hockey. The paper looked at data such as: forward movement / time, forward movement at certain speed / time, forward movement at certain speed in neutral zone / time, and alternate zone entries / time, but was unable to find any correlation with shots on net (Silva et al., 2018).

A Novel Approach to Quantifying Pace in Real-Time Sports: Pace in Hockey

The similarity between all these investigations for pace in real-time sports, is that ideas of pace are currently limited to location-based data of players and/or ball. This data is too expensive/inaccessible to many hobby sports analysts, and is also nearly impossible to track through watching or through manipulating existing data. For this report, we chose to analyse NHL data from the 2023-24 season, due to our mutual interest, and as many people do, we prefer to watch the offensive side of the game vs the defensive side. Therefore, we will explore variables relating to shots on net to explore pace in hockey, using accessible data, to see if there is any relationship between performance and pace.

Data Acquisition and Cleaning

As mentioned previously, while location based metrics seem the most promising thus far in understanding pace of play in real-time sports, these results have only been positive in a soccer application, not hockey. Furthermore, the vast majority of data in that area is inaccessible to us, and the data in that realm that is freely available online (such as NHL edge), is too general for the scope of our project. Therefore, we decided to take an approach of brainstorming 'unique' data points in an attempt to find relationships with shots on goal.

NHL API Data

We decided to look at faceoff percentages, in the offensive, neutral, and defensive zone, as we assumed that a team with a high faceoff percentage in the offensive zone would also get more shots on net, due to securing possession. That assumption was similarly applied for teams poor at defensive zone faceoffs, who would theoretically give up more shots. Furthermore, we looked at fenwick (all unblocked shot attempts) and corsi (all blocked shot attempts + unblocked shot attempts). For the faceoff, fenwick, and corsi data, the data was scraped through the 2 api scrape files found in the nhlAPI subfolder in stage1 and stage2 folder. One file scraped all 5on5 data (when both teams have 5 skaters excluding goalie on the ice), and the other file scraped all situations (5 on 5, 4 on 5, 3 on 5, etc). We chose the 5on5 data for our analysis as in all situations the team with a power play (5 on 4) will of course have more shots and wouldn't represent our problem well, whereas 5on5 keeps it even and we are able to see the true pace for each team.

MoneyPuck Data

We also decided to examine shot data, as we assumed that teams taking certain shots would get more shots on net. Beyond 'standard' shots such as wrist shots, slap shots, et cetera, we also looked at ratios of defender shots/all team shots, and long shots (shots over 41.34 feet, calculated as the length from the goal line at the net to the top of a face-off circle). The last two data points were chosen because long shots are considered 'low danger shots', meaning that they have a low chance of scoring, and consequently a good chance of giving up possession. Thus, if a team were to take a lot of shots from distance, we assume it would be a team which takes shots quickly on offence without taking time to set up, and that they do not mind losing possession in exchange for getting shots on net, thus a pacey team. The shot data was available through MoneyPuck in csv format. The file included a row for every single shot taken during the 2023-24 regular and playoff season, with attributes for each shot such as the gameID, player name, type of shot, distance from net, et cetera. The data was cleaned by sorting for regular season shots only during 5on5 play, along with the calculation of season average per team.

Data Analysis

Correlation Analysis

After gathering this data from various sources, we wanted to find out which categories actually correlated to more shots on net, to construct our idea of 'pace'. For each group of data collected (data from NHL API, and MoneyPuck), each respective group had its own tracking for shots on goal, which resulted in slight discrepancies between the two sets of 'shots on goal'. Thus, the faceoff variables (offensive zone percentage, defensive, neutral) and fenwick/corsi, were compared to the shots on goal collected from the same sources, and likewise for the MoneyPuck shot data. For each group of data, the correlation between the target variable (total shots on net), and the variables of interest, such as wrist shots or defensive zone faceoff percentage, was calculated for a team. The alpha level we chose was 0.05 for significant relationships. We found that every p-value was significant, but the levels of correlation varied. Thus, we decided to test thresholds of r-value for each ml model: 0.0, 0.2, 0.4, and 0.6, which altered the number of columns per r-value.

Machine Learning Analysis

We wanted to group together similar teams in terms of pace and check to see if there is any link with team success. Given we wanted to be able to visualize the data as well we moved forward with PCA which allowed us to retain all important information as well. The data was also scaled using MinMaxScaler() as the ranges for each column varied. To model the data KMeans, Agglomerative, and Affinity clustering was used and we compared the results across each clustering model. These are seen in the stage4-results folder and below in the paper in the visualizations section with each team labeled with their corresponding points from that season.

Results

When analyzing our 3 different clustering outputs we used our prior knowledge of hockey analytics to assist us in our analysis which is a reason as to why knowledge of a given data domain is important. Based on our background, we chose Agglomerative (figures 1 through 4) as the best model due to a couple reasons. One, the model clustered together EDM, CAR, LAK, and FLA consistently in ¾ of the r-value thresholds. These 4 teams were known to be the best at controlling the pace of the game and changing it as the game situations changed. Two, the model clustered together varying batches of elite (offensive) teams in the top right of the graph, which seemed to generally pass the eye test. In the first graph, for r > 0.0, this model incorporated every data category we had gathered. Surprisingly, this turned out to be the 'cleanest' out of the four, as we expected a model solely using variables with a high correlation with high shots on goal to differentiate teams more. However, it seems that taking a more holistic approach to which data is used seems better to quantify pace, or at least to our eye it seemed that the graph seemed to generally pass the eye test, with the 4 teams aforementioned clustered together, with EDM and CAR as clear outliers which matches their status as the best offensive teams of the year. There are some teams that do not match other clustered teams in point totals, which can mostly be explained by their defence/goaltending. For example, although NJD only had 81 points, they are positioned among elite offensive teams in the first and second figures. This is because they were victims of "bad luck", as their goaltending was very poor which is not taken into this model, while their forwards and defencemen were performing to the calibre of the rest of the teams in the top right cluster. For figure 2, which utilises a threshold r value of 0.6 between the target variable and shots on net (which limited the number of categories from figure 1), CAR and EDM further distance themselves from the other teams, which matches up with the eye test of many observers last season which classified these two teams as possessing an offence a tier above other teams. Overall, we believe that the Agglomerative clustering model was able to capture pace of play for different teams and showcased a relationship between the pace a team plays at and their success.

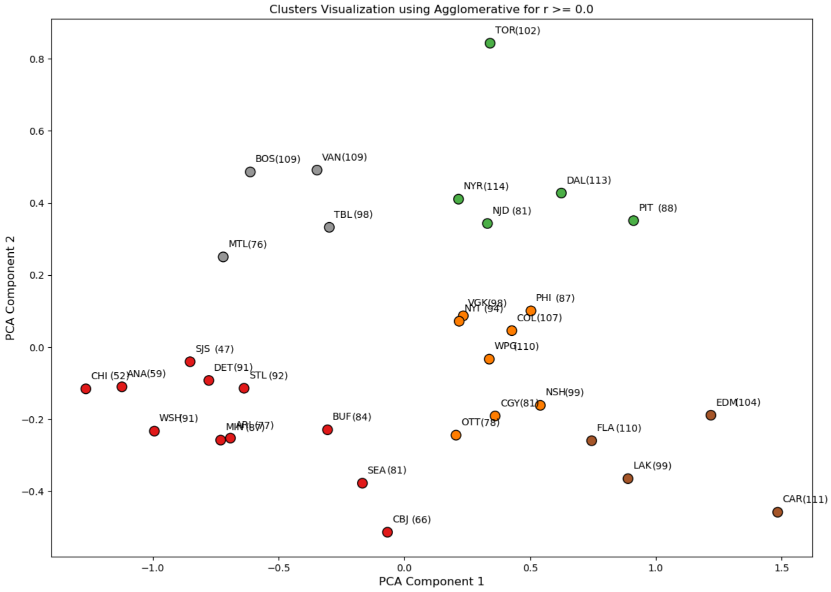

Figure 1.

Agglomerative Clustering using variables with r > 0 (every variable)

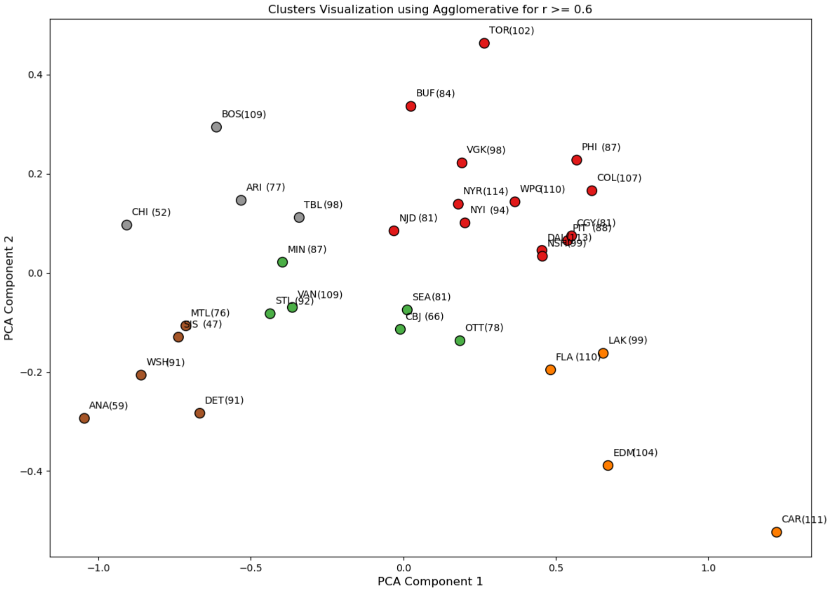

Figure 2.

Agglomerative Clustering using variables with r > 0.6 (less variables)

References

Dona, N. E., & Swartz, T. (2023). A causal investigation of pace of play in soccer. *Statistica Applicata - Italian Journal of Applied Statistics*, *35*(1). https://doi.org/10.26398/IJAS.0035-006

Silva, R. M., Davis, J., & Swartz, T. B. (2018). The evaluation of pace of play in hockey. *Journal of Sports Analytics*, *4*(2), 145–151. https://doi.org/10.3233/jsa-170192

Project Experience Summary

Dilpreet Grewal

In the collection of the data I scraped the NHL Api to get all game by game shot data and faceoff data from the 2023-24 NHL season. I used the .get() and pandas functions to obtain the specific data I needed and further cleaned the data to the required format. Using the game by game data I created a dataframe that had season averages for all teams by indexing the required columns and performing calculations using functions such as .sum(). The collection and cleaning of the data allowed us to perform further statistical analysis and machine learning in which we needed season average data for each team.

In the machine learning section I clustered together similar teams based on their pace of play to check if there is any correlation between pace and success. In order to do this I first used Principal Component Analysis with 2 components which would allow us to visualize the data. Then, the dataset was scaled using MinMaxScaler() as there was a wide array of ranges. Lastly, I performed 3 clustering algorithms on the data. From the results, we found that most of the successful teams were being clustered together and the same was said for most of the unsuccessful teams. Aside from a few outliers, we concluded that there was a relationship between the pace of a team's play and their success.

Justin Mahun

Wrote and formatted a technical report, simplifying a wide range of data science concepts for those with a basic understanding of statistics to communicate results.

Identified and scraped relevant data sources to enhance machine learning model performance, leveraging pandas to process and clean data, which improved the clustering model accuracy.

Conducted statistical analysis using pandas to identify key data subsets to model relationships between key variables, which uncovered insights to improved machine learning models.

Reorganized a GitHub repository to improve clarity and usability by restructuring files and directories and creating a detailed README. Documented the purpose of each folder, provided step-by-step instructions for running the project, and automated tasks such as file execution and library installation for streamlining project running for new users.