Lessons, Observations and Mistakes from 10 years of Analyzing 'Found Hockey Data'

Michael Schuckers

St. Lawrence Univ.
Statistical Sports Consulting, LLC
schuckers@stlawu.edu
@SchuckersM
@EmpiricalSports

March 3, 2018
In the days after CASSIS in early Fall 2016 I said: I don’t know who will win the US Election, but I will bet they would be impeached within a year.
In the days after CASSIS in early Fall 2016 I said:
In the days after CASSIS in early Fall 2016 I said:

*I don’t know who will win the US Election, but I will bet they would be impeached within a year.*
Somewhat Relevant Biases

- Tenured academic, primarily teaching institution
- PhD Statistician, Bayesian
- Played HS football in Penn., college club soccer
- Coach summer travel soccer - U12, U15 Boys & Coed
- American living 90 min from Lake Placid (6 rinks in 20 min)
- 6+ concussions (none long term unlike my kids)
- Learned R when it was still S
- Analytically hate the Bettman point & Shootout

https://groups.yahoo.com/neo/groups/HAG_list/info

‘The Hockey Analysis Group (HAG) is home to fact-based analysis and discussion of hockey.’

First couple of months saved posts by:
R. Vollman, D. Johnson, T. Awad, A. Ryder, K. Krzywicki, T. Tango …
Shootout or Crapshoot

First public work on the shootout (poster), NCCORS (2010):

Shootout or Crapshoot: An Analysis of the NHL Shootout after Five Years
Michael Schuckers
St. Lawrence University and Statistical Sports Consulting

Model
Following Albert and Cohn (1999), let
\[ y_i \sim \text{Bernoulli}(\pi_i) \]
for the $i$th shootout shot with
\[ \pi_i = \Phi(\beta_0 + \beta_1 x_i) \]
Where $\beta_0$ is the overall mean, $\beta_1$ is the shooter effect ($x_i = 1$ for top shooters, $x_i = -1$ for bottom shooters) and $\Phi$ is the cumulative Gaussian/Normal distribution

Prior distribution ($\beta_0, \beta_1 \sim N(0,100)$)

Analysis
Posterior generated using MCMC (from MCMCpack in R).
4 chains of 50000 (taking every 50th iteration)
Burn-in of 1000 iterations
Convergence for each parameter using Gelman & Rubin criteria

Results
Among players with more than 10 shots, 99% CI’s all $\beta_0$ and $\beta_1$ contain 0 except

Goals
Marc Denis 6/11(54.5%) - Most goals
Shooters
Michael Frolik 3/5 (60%)
Dany Heatley 3/3 (100%)
Tomas Plekanec 2/3 (66.6%)
Taylor Pyatt 2/2 (100%)
Bobby Ryan 1/1 (100%)
Michael Ryder 1/2 (50%)
Stephan Weiss 0/3 (0%)

Conclusions/Next Steps
- Mostly CRAPSHOOT
- Evidence that some shooters are worse than league average
- No evidence that some shooters are better than league average
- Confidence that $N(0,100)$ very flat prior heavily assume player differences
- Better model: full hierarchical Bayesian model with terms for average goalie and average shooter effect
- NHL rule change for 2010-11
- Tiebreaker for regular season standings
- No longer includes Shootout Wins
- Only regulation and overtime wins

Thanks to Jessica Chapman (SLU) and Chris Wells (SLU & SSC) for comments
Some Mistakes

Shootout or Crapshoot

Bayesian Probit model with mean + shooter + goalie (simple model)
Data from 2005-06 to 2009-10
Conclusions:

- Mostly CRAPSHOOT
- 112 goalies, 99% CI, 1 significant
- 571 shooters, 99% CI, 10 significant (all low)

‘All models are wrong. Some are useful.’ - George Box

Michael Schuckers (St. Lawrence Univ. Statistical Sports Consulting, LLC
schuckers@stlawu.edu @SchuckersM @EmpiricalSports

VANHAC-Schuckers
March 3, 2018
Shootout or Crapshoot

Bayesian Probit model with mean + shooter + goalie (simple model)
Data from 2005-06 to 2009-10
Conclusions:

- Mostly CRAPSHOOT
- 112 goalies, 99% CI, 1 significant
- 571 shooters, 99% CI, 10 significant (all low)

‘All models are wrong. Some are useful.’ - George Box
Shootout or Crapshoot

Bayesian Probit model with mean + shooter + goalie (simple model)
Data from 2005-06 to 2009-10
Conclusions:

- Mostly CRAPSHOOT
- 112 goalies, 99% CI, 1 significant
- 571 shooters, 99% CI, 10 significant (all low)

‘All models are wrong. Some are useful.’ - George Box
Model too simple. Multiplicity? Missing other factors (Covariates)
Another analysis (in 2013) about who goes first (asked for for free by NHL team).
Data through 2011-12.
Simple summaries:
Another analysis (in 2013) about who goes first (asked for for free by NHL team).
Data through 2011-12.
Simple summaries:

- Away teams won slightly more (52.6%, p<0.05)
- Away teams go first more often (67.0%, p<0.01)
- Going first results in winning (49.5%, p>0.30)
More Shootout

Another analysis (in 2013) about who goes first (asked for for free by NHL team).
Data through 2011-12.
Simple summaries:

- Away teams won slightly more (52.6%, p<0.05)
- Away teams go first more often (67.0%, p<0.01)
- Going first results in winning (49.5%, p>0.30)

Could have been better: Controlling for Covariates...
Another analysis (in 2013) about who goes first (asked for for free by NHL team).
Data through 2011-12.
Simple summaries:

- Away teams won slightly more (52.6%, $p < 0.05$)
- Away teams go first more often (67.0%, $p < 0.01$)
- Going first results in winning (49.5%, $p > 0.30$)

Could have been better: Controlling for Covariates... and it was free.
Shootout: Found Data

Remedy all this in *Lopez and Schuckers* (2017), actual paper.

Bayesian Logit model a la Albert & Chib

- Goalie
- Shooter
- Status (Win Imminent, Loss Imminent)
- Home or Visitor
- Defencemen or Forward

Takeaway

I WAS WRONG. Shootout involves more differentiation of skill that initial analysis.

We improved our approach & used additional variables.

Still some issues including incorporating coaching knowledge a la Lock & Nettleton.

And I still had it analytically but . . .
Takeaway

I WAS WRONG. Shootout involves more differentiation of skill that initial analysis.

We improved our approach & used additional variables.

Still some issues including incorporating coaching knowledge a la Lock & Nettleton.

And I still had it analytically but . . .
Goaltending - DIGR

Next set of mistakes:

DIGR - Defense Independent Goalie Rating, Schuckers (2011)
MIT Sloan Research Paper Finalist

Adjusted SV% based upon location (x,y), shot type & strength
Top Left: M Brodeur (NJD) Slap at ES
Top Right: T Thomas (BOS) Wrist at PK
Bottom Left: M-A Fleury (PIT) Snap at ES
Bottom Right: I Bryzgalov (PHX) Slap at PK
Innovation:

- Spatial smoothing (LOWESS) for each player
- Different surface for each shot type (Wrist, Slap, etc.)
- Specify math framework for adj SV%
- Adj rel to avg shot dist$^{bn}$ OR rel to avg goalie
Collective response from Hockey Analytics . . .
Collective response from Hockey Analytics . . . Meh
Collective response from Hockey Analytics . . . Meh

- I didn’t look at reliability (odd v. even)
- I didn’t look at repeatability (out of sample)
- and those pesky covariates (screens)
Collective response from Hockey Analytics . . . Meh

- I didn’t look at reliability (odd v. even)
- I didn’t look at repeatability (out of sample)
- and those pesky covariates (screens)

I SCREWED UP, so a nice method was largely ignored.
Did this important work and added Bayesian shrinkage component \((n^\dagger)\)  
Albert et al (eds.), (2016)

**Table 1:** Intraseason correlation of DIGR for even and odd shots using \(n^\dagger = 1000\)

<table>
<thead>
<tr>
<th>Season</th>
<th>More than 500 shots faced</th>
<th>More than 750 shots faced</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Correlation</td>
<td>(J)</td>
</tr>
<tr>
<td>2009/10</td>
<td>0.288</td>
<td>48</td>
</tr>
<tr>
<td>2010/11</td>
<td>0.644</td>
<td>45</td>
</tr>
<tr>
<td>2011/12</td>
<td>0.655</td>
<td>45</td>
</tr>
<tr>
<td>2012/13</td>
<td>0.251</td>
<td>24</td>
</tr>
</tbody>
</table>
Table 2: Correlation between a goalie's DIGR in one year and their DIGR in subsequent years.

<table>
<thead>
<tr>
<th>Seasons</th>
<th>More than 500 shots</th>
<th></th>
<th>More than 750 shots</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Correlation</td>
<td>J</td>
<td>Correlation</td>
<td>J</td>
</tr>
<tr>
<td>2009/10 v 2010/11</td>
<td>0.237</td>
<td>34</td>
<td>0.393</td>
<td>22</td>
</tr>
<tr>
<td>2010/11 v 2011/12</td>
<td>0.647</td>
<td>37</td>
<td>0.665</td>
<td>28</td>
</tr>
<tr>
<td>2011/12 v 2012/13</td>
<td>-0.174</td>
<td>15</td>
<td>0.677</td>
<td>6</td>
</tr>
</tbody>
</table>
'Found Hockey Data' borrowing from Nick Horton (Amherst College)

Statistical/Mathematical Approach

NHL Hockey is 40+ dimensional process
We model it (or drop data) to deal with imbalance.
Correlational Data

Causation $\rightarrow$ Outcome

Examples:
- B. Horvat $\rightarrow$ Higher xG
- 4F,1D $\rightarrow$ Better PP
- Carry puck into OZ $\rightarrow$ Higher CF%
Correlational Data

Cause → Outcome

Examples:

B. Horvat → Higher xG
Correlational Data

$\text{Cause} \rightarrow \text{Outcome}$

Examples:

B. Horvat $\rightarrow$ Higher xG

4F,1D $\rightarrow$ Better PP
Correlational Data

*Cause* → *Outcome*

Examples:

- B. Horvat → Higher xG
- 4F,1D → Better PP
- Carry puck into OZ → Higher CF%
Correlational Data

*Cause → Outcome*

Examples:

B. Horvat → Higher xG

4F,1D → Better PP

Carry puck into OZ → Higher CF%

Assume *Ceteris Paribus* - all else being equal
Ceteris Non Paribus is the norm for ‘found’ data not designed data. (Statistical Design of Data Collection)
Need to use methods that account for other factors
Zone Starts

Players

Corsi Rel For %
Found Data

Diagram:

- **Players**
- **QoT**
- **Corsi Rel For %**
Suppose we study G’s in Wild Illustration.
Suppose we study G’s in

Enlist an army of recorders:
Suppose we study G’s in

Enlist an army of recorders:

Record every time G has choice:
  Play puck or leave for D
Suppose we study G’s in

Enlist an army of recorders:

Record every time G has choice:
  Play puck or leave for D

Record whether subsequent Zone Entry was Successful
Suppose we study G’s in 

Enlist an army of recorders:

Record every time G has choice: 
Play puck or leave for D

Record whether subsequent Zone Entry was Successful

Determine: Play puck in Trap $\xrightarrow{?} HigherZoneEntry\%$
Example 1: A. Vasilevsky (TBL)
Choices
- Play puck
Example 1: A. Vasilevsky (TBL)
Choices
- Play puck
- Leave for V. Hedman
Example 1: A. Vasilevsky (TBL)
Choices
- Play puck
- Leave for V. Hedman
- Leave for A. Stralman
Example 1: A. Vasilevsky (TBL)
Choices
- Play puck
- Leave for V. Hedman
- Leave for A. Stralman
- Leave for M. Sergachev
Example 1: A. Vasilevsky (TBL)
Choices

- Play puck
- Leave for V. Hedman
- Leave for A. Stralman
- Leave for M. Sergachev
- Leave for B. Coburn
Example 1: A. Vasilevsky (TBL)
Choices

- Play puck
- Leave for V. Hedman
- Leave for A. Stralman
- Leave for M. Sergachev
- Leave for B. Coburn
- Leave for R. McDonagh
Example 1: A. Vasilevsky (TBL)
Choices

- Play puck
- Leave for V. Hedman
- Leave for A. Stralman
- Leave for M. Sergachev
- Leave for B. Coburn
- Leave for R. McDonagh
- Leave for D. Girardi
Example 2: J. Quick (LAK) Choices

- Play puck in
- Leave for D. Doughty
Example 2: J. Quick (LAK)

Choices

- Play puck in
- Leave for D. Doughty
- Leave for J. Muzzin
Example 2: J. Quick (LAK)

Choices

- Play puck in
- Leave for D. Doughty
- Leave for J. Muzzin
- Leave for A. Martinez
Example 2: J. Quick (LAK)
Choices
- Play puck in
- Leave for D. Doughty
- Leave for J. Muzzin
- Leave for A. Martinez
- Leave for D. Forbort
Example 2: J. Quick (LAK)

Choices

- Play puck in
- Leave for D. Doughty
- Leave for J. Muzzin
- Leave for A. Martinez
- Leave for D. Forbort
- Leave for C. Folin
Example 2: J. Quick (LAK)

Choices

- Play puck in
- Leave for D. Doughty
- Leave for J. Muzzin
- Leave for A. Martinez
- Leave for D. Forbort
- Leave for C. Folin
- Leave for D. Phaneuf

Perhaps Quick behaves differently than Vasilevsky
In a non-Found Data world (Statistical Design of Experiments)

- Every goalie try Play Puck/Leave in same ratio
- Do so equal ratio with every combination of players
- Every opponent
- Every game situation
- Some kind of random order
Impact of teammates is unequal on goalies because data is found.

Better account for those teammates (D specifically?) on the ice as part of my analysis. Why we have regression.
Deal with found data

1. Drop data (only close 5v5)
2. Model data (Regression & Smoothing)

Regression Adjusted Plus-Minus approach B. Macdonald via Ilardi and Barzalai
Scale of Data Analysis

The scale matters for data analysis.

One tenet of Statistics, the discipline, is that your unit of analysis should be the largest unit that is impacted by your factors. In the case of sports that often means plays or shifts or possessions. Consequently, doing analysis at a game or season level is not ideal since it doesn’t disentangle all of the important impacts.
Scale of Data Analysis

Impact of choosing wrong

- TOO BIG → average out effects (e.g. QoC)
- TOO SMALL → pseudoreplication,
Hockey Analysis Scale

Teams: Game (not shot)
Players: Shift or Events
Draft: Player
Zone Entry: Entry

Get scale right and measure as many variables as you can
Link to 2012 Panel & Faceoffs
https://www.youtube.com/watch?v=CF_u6wDfc_w&t=29m23s
NHL Data is

Fine or Fixable in Aggregate

Lots of old-school hand wringing
  MSG, in particular

Lots to not like about NHL.com RTSS data

But ... pretty good on aggregate
Rink Effects

David Perron: ‘Even NHL stats, when they do Hits or things like that, it’s not consistent from building to building.’
Rink Effects

David Perron: ‘Even NHL stats, when they do Hits or things like that, it’s not consistent from building to building.’

TRUE for HITS and some other stuff but not everything
Rink Effects

David Perron: ‘Even NHL stats, when they do Hits or things like that, it’s not consistent from building to building.’

TRUE for HITS and some other stuff but not everything

But we can adjust for (relative) rink differences again assuming patterns of recording shot up over large amounts of data.

http://www.hockeybuzz.com/blog/Sheng-Peng/
David-Perron-on-Why-Players-Resist-Analytics--Player-Tracking/244/90978
Rink Effects
Rink Effect: Fixes

- Schuckers & Curro (2013) proposed simple adjustment to X- & Y-coordinates of shots
- Schuckers & Macdonald (2014) proposed count adjustments for events
- Improve on x- & y- for NHL by using image warping * & Schuckers et al. (20??) — NEED A COAUTHOR
NHL Data: Rink Effects

Schuckers & Macdonald (2015)

Table 3: Yearly Coefficients for SHOTs.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>mean, exp((\hat{\mu}))</td>
<td>23.637</td>
<td>24.935</td>
<td>25.313</td>
<td>26.448</td>
<td>25.616</td>
<td>25.089</td>
</tr>
<tr>
<td>ASD, exp((\hat{\beta}))</td>
<td>0.964</td>
<td>0.967</td>
<td>0.955</td>
<td>0.957</td>
<td>0.955</td>
<td>0.957</td>
</tr>
<tr>
<td>home ice, exp((\hat{\gamma}))</td>
<td>1.077</td>
<td>1.061</td>
<td>1.051</td>
<td>1.054</td>
<td>1.062</td>
<td>1.033</td>
</tr>
</tbody>
</table>

ASD is average score difference
NHL Data: Rink Effects

Schuckers & Macdonald (2015)

Table 4: Significant Rink Effects for SHOTs.

<table>
<thead>
<tr>
<th>Rink</th>
<th>Effect</th>
</tr>
</thead>
<tbody>
<tr>
<td>FLA</td>
<td>1.030</td>
</tr>
<tr>
<td>STL</td>
<td>0.955</td>
</tr>
</tbody>
</table>

There were no persistent 'homer' effects for SHOTs.

'Significant' here is consistent (5/6 yrs), all sign in same direction & sign in all years model
Selection of NHL players can be improved by using analytics PERIOD.

- Draft Value Charts are available (Tulsky, Schuckers, others)
- NHL drafting is better than Central Scouting Ranking (Schuckers & Argeris, 2015)
- NHL drafting can be improved with analytics (Jessop, Weissbock, Lawrence)
  - Sham Sharron (Weissbock, Jessop, 2014)
  - Prospect Cohort Success (Weissbock, Lawrence, Hohl, Jessop, various)
  - Draft by Numbers (Schuckers, 2016)
  - Text Mining Scouting Reports (Seppa, Schuckers & Rovito, 2017)
- Survival Analysis of NHL Prospect Timelines (Nandakumar, 2017)
Draft & ROI on Analytics

Wake of S. Mehta (and ‘Summer of Analytics’) being hired in NJ, wrote with S. Argeris (now Carolina Panthers) about value of analytics:
For drafting, I said

\[ \text{Conservatively the total for drafting is about $895K per year from drafting. This excludes something like the creation and use of a value pick chart.} \]

AND you build depth and resources.
See, Vegas & T. Tatar.
Draft & ROI on Analytics

Wake of S. Mehta (and ‘Summer of Analytics’) being hired in NJ, wrote with S. Argeris (now Carolina Panthers) about value of analytics:

For drafting, I said

> [C]onservatively the total for drafting is about $895K per year from drafting. This excludes something like the creation and use of a value pick chart.

AND you build depth and resources.

See, Vegas & T. Tatar.

See, Yzerman, Steve & TBL.

http://statsportsconsulting.com/2014/05/06/
return-on-investment-for-hockey-analytics-1/
“We will stick with our simple guideline: The less the prudence with which others conduct their affairs, the greater the prudence with which we must conduct our own.” - Warren Buffett

Health of Hockey Analytics
Health of Hockey Analytics

Better than ever
Health of Hockey Analytics

Better than ever

- HAC’s (Rob Vollman, RITHAC, OTTHAC, VANHAC, etc.)
Health of Hockey Analytics

Better than ever

- HAC’s (Rob Vollman, RITHAC, OTTHAC, VANHAC, etc.)
- Mentoring: s/o to Hockey-graphs.com
Health of Hockey Analytics

Better than ever

- HAC’s (Rob Vollman, RITHAC, OTTHAC, VANHAC, etc.)
- Mentoring: s/o to Hockey-graphs.com
- GITHUB’s, More & more open data, open code
Health of Hockey Analytics

Better than ever

- HAC’s (Rob Vollman, RITHAC, OTTHAC, VANHAC, etc.)
- Mentoring: s/o to Hockey-graphs.com
- GITHUB’s, More & more open data, open code
- Tracking Data (someday?): Why but not what
Health of Hockey Analytics

Better than ever

- HAC’s (Rob Vollman, RITHAC, OTTHAC, VANHAC, etc.)
- Mentoring: s/o to Hockey-graphs.com
- GITHUB’s, More & more open data, open code
- Tracking Data (someday?): Why but not what
- Crowdsourcing (Tape-to-Tape, Passing Project, All Three Zones project, NWHL Big Stat Track, etc.)
Health of Hockey Analytics

Better than ever

- HAC’s (Rob Vollman, RITHAC, OTTHAC, VANHAC, etc.)
- Mentoring: s/o to Hockey-graphs.com
- GITHUB’s, More & more open data, open code
- Tracking Data (someday?): Why but not what
- Crowdsourcing (Tape-to-Tape, Passing Project, All Three Zones project, NWHL Big Stat Track, etc.)
- Paths more open, opportunities are greater
Health of Hockey Analytics

Better than ever

- HAC’s (Rob Vollman, RITHAC, OTTHAC, VANHAC, etc.)
- Mentoring: s/o to Hockey-graphs.com
- GITHUB’s, More & more open data, open code
- Tracking Data (someday?): Why but not what
- Crowdsourcing (Tape-to-Tape, Passing Project, All Three Zones project, NWHL Big Stat Track, etc.)
- Paths more open, opportunities are greater
- Field is healthier PERIOD
Recap

10 years goes fast and not possible with great collaborators
Recap

10 years goes fast and not possible with great collaborators
Including:
D. Lock, C. Knickerbocker,
C. Wells,
M. Generous, L. Daley
J. Hurlbut, S. Bell, D. Driscoll
S. Mills
S. Argeris, T. Seppa, M. Rovito
M. Lopez
J. Sylvain, J. Tank, M. Bost-Brown
G. White, G. Gilman
Recap

- Do stuff but remember the fundamentals (reliability, validity, relevance to winning)
- Don’t be afraid to make mistakes but own up to them
- Hockey data is ‘Found’ Data
- Need to use tools (like RAPM) to deal with correlations/causal implications
- Find the right scale of data analysis
- NHL RTSS Data is ‘fixable’
- More work on Entry Draft
- Hockey Analytics/Statistics is a healthy community
- ...keep working to ensure it gets better
2018 OTTHAC
Sept 14 (social and workshops) & 15,
Carleton University, Ottawa, ON
Thanks
schuckers@stlawu.edu
Tyler Kennedy:
Tyler Kennedy:
Feature not a bug.