Investigating Major League Baseball Amateur Draft Patterns Pertaining to Climate

Searching for proof of the "Cold Weather Bias" in the MLB Draft

By Michael Dudkin—MGIS Student, Penn State

Mike Trout
Millville, NJ
19th Pick, 2009 Draft
59.6 Career WAR

Should your team be wary of the next?
Or excited to unearth the next?

Chris Lubanski
Norristown, PA
5th Pick, 2003 Draft
0 Career WAR

RESEARCH QUESTION

In Major League Baseball, about **75% of the player pool is composed of players acquired through the amateur draft**, held every June. The draft is the single most important amateur talent pipeline to MLB, and the core of each team's talent acquisition strategy, supplementing talent from the Dominican Republic, Cuba, and other places where talent can be signed outside the confines of the draft process.

One common perception in the baseball community is that amateur - particularly high school - **draftees from "cold weather" places are less likely to develop into major leaguers, due to only being able to play in the spring/summer versus year-round for their counterparts in "warm weather" locations such as Florida, California, Texas, and Arizona**. This "**cold weather bias**" may affect the draft position of prospects from Northeastern and other cold weather states. The most prominent **recent example of this phenomenon is Mike Trout** of the Los Angeles Angels who was a highly regarded prospect in high school, but dropped all the way to 25th in the 1st round of the 2009 draft due in part to concerns that he played at a cold-weather high school in Millville, NJ. Trout thereafter became a superstar and several commentators pointed that his cold-weather school, in retrospect, contributed to him falling in the draft. The goal of this study is to investigate if there a cold weather effect in the draft positions of prospects, and **if cold-weather state prospects are systemically undervalued by teams.**

Can a Juco player who is unknown make a name for himself with a great season or do they suffer from not having great competition?

SN Klaw (2:02 PM)

In the south and west, yes, they'll get on the radar. Might be harder in cold-weather areas because they start too late and by the time they get traction it's too late for them to get cross-checked. But in states like FL, AZ, CA, TX, even the deep south, I don't think those players can go unnoticed.

Ask for an explanation of how Trout could still be available after 23 picks, and what you get sounds like a bunch of excuses:

- The weather is too unpredictable, so it's too hard for scouts to plan trips.
- The weather is too cold and wet, so the player's body of work is limited compared to players in other parts of the country.
- Pitchers from the northeast have historically done OK in pro ball, but position players have not.
- Due to shorter seasons, some teams don't scout the area with any regularity.

To a scouting director, those aren't excuses. They're realities of the job.

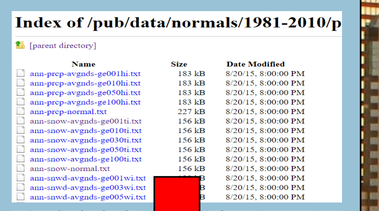
TB 13 OF Garrett Whitley, Niskayuna HS (N.Y.)

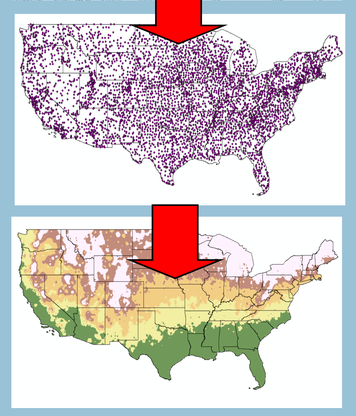
Analysis: Was said to be a candidate for the Diamondbacks as the first overall choice. Part of the reason interest in Whitley has been delayed is because he's from New York, and sometimes players from cold weather states aren't as widely scouted. An example: Mike Trout, who went in the first round but not until the 25th overall pick in 2009. Whitley has great bat speed and leg speed that is just as imposing, frequently beating out otherwise routine infield grounders. He's seen as a center fielder in the majors. Whitley batted .356, going 21 for 59, with three home runs, three doubles, a triple and 13 walks during his senior year. He's also hit a ball that was measured by a scout at 497 feet

Above: Examples of common "cold weather bias" lingo in MLB scouting community

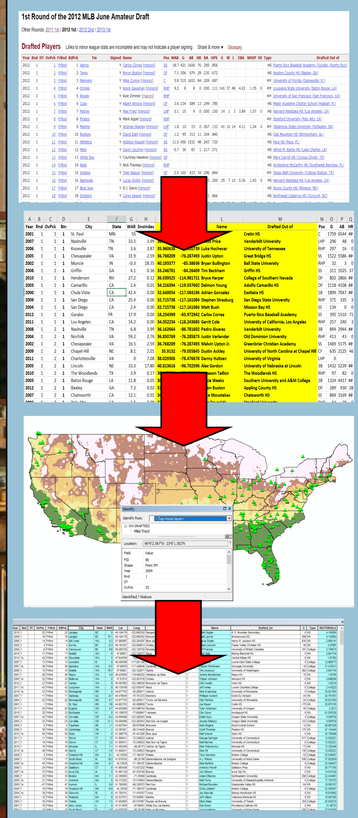
METHODOLOGY

Step 1 - Create 2 custom climate surfaces for US Mainland using NOAA 1981-2010 Climate Normals Database: based on average winter temperature, and average # of snowy days/year; create 5 climate categories

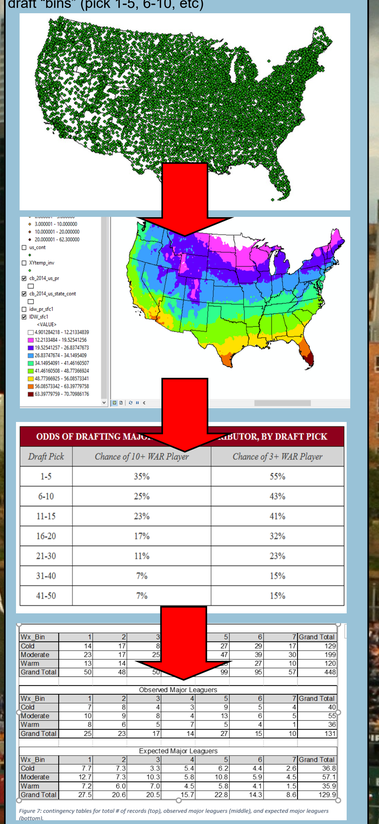




Step 2 - Obtain Baseball-Reference.com data on 1st Round Draftees 2000-2012; geocode hometowns of draftees' hometowns; plot each player on the climate surfaces; assign raster value to each data point



Step 3 - Perform phase 1 analysis on draftees; Assign simple cold/moderate/warm categories to draftees based on hometown's avg winter temp; apply Matthew Murphy's (2014) probabilities of 3+ WAR and create draft "bins" (pick 1-5, 6-10, etc)



ODDS OF DRAFTING MA… …BUTOR, BY DRAFT PICK		
Draft Pick	*Chance of 10+ WAR Player*	*Chance of 3+ WAR Player*
1-5	35%	55%
6-10	25%	43%
11-15	23%	41%
16-20	17%	32%
21-30	11%	23%
31-40	7%	15%
41-50	7%	15%

Wx_Bin	1	2	3	4	5	6	7	Grand Total
Cold	14	17	27	29	27	29	17	129
Moderate	23	17	25	33	47	39	30	199
Warm	13	14	24	37	25	27	10	120
Grand Total	50	48	76	99	99	95	57	448

Observed Major Leaguers

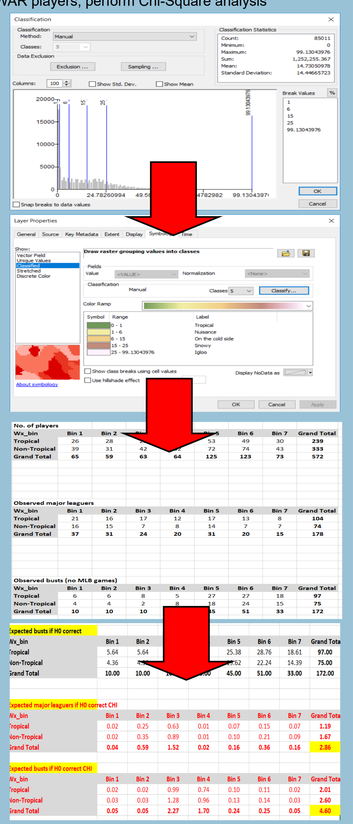
Wx_Bin	1	2	3	4	5	6	7	Grand Total
Cold	7	8	4	3	9	5	4	40
Moderate	10	9	8	4	13	6	5	55
Warm	8	6	5	7	5	4	1	36
Grand Total	25	23	17	14	27	15	10	131

Expected Major Leaguers

Wx_Bin	1	2	3	4	5	6	7	Grand Total
Cold	7.7	7.3	3.3	5.4	6.2	4.4	2.6	36.8
Moderate	12.7	7.3	10.3	5.8	10.8	5.9	4.5	57.1
Warm	7.2	6.0	7.8	4.5	5.8	4.1	1.5	35.9
Grand Total	27.5	20.6	20.5	15.7	22.8	14.3	8.6	129.9

Figure 7: contingency tables for total # of records (top), observed major leaguers (middle), and expected major leaguers (bottom)

Step 4 - Parse "average snowy days" data in 2 ways—one granular by creating 5 different "snowiness" categories; and then binary (tropical vs non-tropical climate); create probabilities chart for "busts" in additional to 3+ WAR players; perform Chi-Square analysis



RESULTS

Goal: Calculate a chi-square statistic with which we are able to reject the null hypothesis at the 95% confidence level

Null hypothesis: that for any given range of slots (e.g. picks 1-5, 6-10 and so on) within the first round, the climate associated with a drafted player's school has no bearing on the likelihood that that player will become a viable major league player (3WAR or greater), or be a draft Bust (No MLB games played)

PHASE1:
Drafted players and **average winter temperature** climate surface (3+ WAR; Murphy probabilities)

All Players **χ2**: 8.38
To be exceeded: 21.02
H0 cannot be rejected

HS Players only **χ2**: 12.11 (3+ WAR; Murphy probabilities)
To be exceeded: 21.02
H0 cannot be rejected

PHASE2:
Drafted players and **average number of snow days** climate surface (3+ WAR; 5 categories; Murphy probabilities)
All Players **χ2**: 18.80
To be exceeded: 36.42
H0 cannot be rejected

Drafted players and **average number of snow days** climate surface (3+ WAR; 5 categories; calculated probabilities)
All Players **χ2**: 20.19
To be exceeded: 36.42
H0 cannot be rejected

Drafted players and **average number of snow days** climate surface (Zero ML Games; 5 categories; calculated probabilities)
All Players **χ2**: 29.12
To be exceeded: 36.42
H0 cannot be rejected
****Empirically there are FEWER busts than expected in the colder categories; however the result only exceeds threshold at 75% confidence level; therefore we CANNOT reject assumption that this observation is simply a result of chance.****

Drafted players and **average number of snow days** climate surface (3+ WAR; 2 categories; calculated probabilities)
All Players **χ2**: 2.86
To be exceeded: 12.59
H0 cannot be rejected

Drafted players and **average number of snow days** climate surface (Zero ML Games; 2 categories; calculated probabilities)
All Players **χ2**: 4.60
To be exceeded: 12.59
H0 cannot be rejected

METHODOLOGY NOTES

- The draft data consisted of 572 records, spanning all 1st rounds of MLB drafts from 2000-2012. Using more recent drafts could not yield reliable data on 3+ WAR players as many are yet to make their debuts. These players may still be prospects, and it is often too early to assign them to one of the "bust" or "3+ WAR" categories definitively.
- The data did NOT include duplicate records of players who were drafted once, did not sign, then were drafted again. ONLY records of players who signed were used.
- Data was limited to first round only; while there are certainly cases of All-Stars and other viable major leaguers being chosen in rounds 2, 3, 12, or later (e.g. Albert Pujols), the overwhelming majority of post-1st rounders do not make the majors, and moreover there is tremendous variability WITHIN the first round. The expected WAR of draft picks plummets after the first handful of picks, and the curve flattens out considerably through the end of Round 1, the supplemental picks, and then the rest of the draft.

DATA NOTES (2)

- A full 333 out of 572 draft records were in the "Tropical" category; this reflects the scouting truism that the (perceived) elite talent resides in the warmer climates of Florida, Texas, Southern California, and the Bay Area.
- In Phase 1, I isolated High School players since a lot of the cold-weather bias commentary concerns Northeastern high schoolers and their alleged lack of practice time, poor competition, and limited seasons.
- In Phase 2 I decided to include all records, since the data revealed there is a distinct lack of high draft picks from large, well-funded cold-weather colleges either. This is partly tautological, as a lot of the elite baseball powerhouse college programs are based in warmer climates.
- E.g., in my database of 1st rounders, there were more draft picks out of the 2 large Los Angeles schools (UCLA, USC) than the universities of Notre Dame, Nebraska, Michigan, Wisconsin, Minnesota, and Washington combined.

DATA NOTES

- The raster surface created from NOAA data had to have breaks between climatic categories that fit the distribution of draft picks—therefore I created breaks in the surface categories on the lower ends of the "number of snowy days per year" scale. A location with 1 or fewer snowy days was considered "Tropical," 1-6 days categorized as "Nuisance," 6-15 days categorized as "On the Cold Side," 15-25 days as "Snowy," and finally 25+ days as "Igloo."
- The reason for choosing this metric over others (e.g. the average temperature metric used in Phase 1) is that it corresponds more to the traditional "cold weather bias" scouting objections—a snowy location not only indicates the presence of cold and lack of ability to play year round, but also limits potential practice and game opportunities in marginal months such as October and April, when weather could create adverse impacts not found better climates
- Because over half the players came from places with virtually no snow, I also re-grouped them into simple binary "Tropical" and "Non-Tropical" categories and then performed same analysis.

TECHNICAL NOTES

- The climate surface analysis in this study was done with the GIS program ArcMap 10.6
- The bulk NOAA climate normal datasets and baseball-reference were processed in Excel, and added as point layers in ArcMap. The draftees' hometowns were converted to lat/long coordinates with widely available online "batch geocoders."
- The probabilities associated with a given bin of draft picks becoming a 3+ WAR major league contributor were taken from research done by the Hardball Times' Matthew Murphy (2014); the "calculated" probabilities using the 2000-2012 data alone, based on marginal totals, yielded similar results - 166 vs 172 expected major leaguers out of 572 based on each respective method.
- Climate surfaces were based on 6000+ data points from NOAA datasets, using the "Inverted Distance Weighed" (IDW) interpolation method, using an exponent of 2, to yield a relatively coarse surface. The nearest 12 neighbor points were used in the interpolation.

SPECIAL ACKNOWLEDGMENETS

This research is possible thanks to support from my Penn State instructors, advisors and administrators. Special thanks to my advisor Eliza Richardson, Justine Blanford, Beth King, Kary Isett, and my class instructors going back to 2014. Thank you to SABR's Scott Fischthal for guiding me in convention planning. A special thank you to my wife Megan and daughter Elise**

EXAMPLE

No. of players

Wx_bin	Bin 1	Bin 2	Bin 3	Bin 4	Bin 5	Bin 6	Bin 7	Grand Total
Igloo	2	1	3	4	6	4	2	22
Snowy	3	2	1	6	3	4	4	23
On the cold side	5	5	6	3	16	18	8	61
Nuisance	16	20	11	19	28	23	16	133
Tropical	39	31	42	32	72	74	43	333
Grand Total	65	59	63	64	125	123	73	572

Observed busts (no MLB games)

Wx_bin	Bin 1	Bin 2	Bin 3	Bin 4	Bin 5	Bin 6	Bin 7	Grand Total
Igloo	0	0	0	1	2	2	1	6
Snowy	0	0	0	3	0	4	7	14
On the	3	1	1	0	5	6	3	19
Nuisance	1	3	1	4	11	12	4	36
Tropical	6	6	8	5	27	27	18	97
Grand To-	10	10	10	13	45	51	33	172

Note handful of categories have fewer than expected by probabilities alone

Expected bust distribution (if H0 cor-

Wx_bin	Bin 1	Bin 2	Bin 3	Bin 4	Bin 5	Bin 6	Bin 7	Grand Total
Igloo	0.35	0.35	0.35	0.45	1.57	1.78	1.15	6.00
Snowy	0.81	0.81	0.81	1.06	3.66	4.15	2.69	14.00
On the	1.10	1.10	1.10	1.44	4.97	5.63	3.65	19.00
Nuisance	2.09	2.09	2.09	2.72	9.42	10.67	6.91	36.00
Tropical	5.64	5.64	5.64	7.33	25.38	28.76	18.61	97.00
Grand To-	10.00	10.00	10.00	13.00	45.00	51.00	33.00	172.00

CHI-Square Values (24 Degrees of Freedom)

Wx_bin	Bin 1	Bin 2	Bin 3	Bin 4	Bin 5	Bin 6	Bin 7	Grand Total
Igloo	0.35	0.35	0.35	0.66	0.12	0.03	0.02	1.87
Snowy	0.81	0.81	0.81	3.56	3.66	0.01	6.93	16.60
On the	3.25	0.01	0.01	1.44	0.00	0.02	0.11	4.85
Nuisance	0.57	0.39	0.57	0.60	0.27	0.16	1.22	3.79
Tropical	0.02	0.02	0.99	0.74	0.10	0.11	0.02	2.01
Grand To-	5.01	1.59	2.73	7.00	4.15	0.33	8.31	29.12

REFERENCES

- Batch Geocoding. (2016). Retrieved from findlatitudeandlongitude.com: http://www.findlatitudeandlongitude.com/batch-geocode/
- Draft Index. (n.d.). Retrieved from Baseball-Reference.com: http://www.baseball-reference.com/draft/
- Murphy, M. (2014, May 22). The Net Value of Draft Picks. Retrieved from The Hardball Times: http://www.hardballtimes.com/the-net-value-of-draft-picks/
- NOAA National Centers for Environmental Information. (2016). Retrieved from 1981-2010 U.S. Climate Normals: https://www.ncdc.noaa.gov/data-access/land-based-station-data/land-based-datasets/climate-normals/1981-2010-normals-data
- Batch Geocoding Tool: http://www.mapdevelopers.com/batch_geocode_tool.php
- NOAA Data Explainer: https://www1.ncdc.noaa.gov/pub/data/normals/1981-2010/readme.txt
- Drafting Extrema: The Expectation of Draft Picks (2014). Retrieved from Beyond the Box Score: https://www.beyondtheboxscore.com/2014/7/31/5950043/mlb-draft-picks-average-expectations-extremes
- Percentage Points of Chi-Square Distribution Table; From the University of Nebraska Lincoln: http://passel.unl.edu/Image/Namuth-CovertDeana956176274/chi-sqaure%20distribution%20table.PNG
- Wasserstrom, Chuck, Inside the Draft Room—The 2009 Anaheim Angels. Retrieved from MLB Trade Rumors:https://www.mlbtraderumors.com/2017/04/inside-the-draft-room-the-2009-angels.html
- ESPN.com SportsNation Chat with Keith Law. Retrieved from http://www.espn.com/sportsnation/chat/_/id/51780/mlb-insider-keith-law
- 2015 MLB First-Year Player Draft: First round complete results. CBS Sports. Retrieved from https://www.cbssports.com/mlb/news/2015-mlb-first-year-player-draft-first-round-complete-results/
- Chris Lubanski. ESPN.com. Retrieved from http://www.espn.com/mlb/player/stats/_/id/29392/chris-lubanski
- Ruth E Hendricks Photography. PNC Park. Retrieved from https://rutheh.files.wordpress.com/2012/05/pncpark.jpg