NBER WORKING PAPER SERIES

NOTHING PROPINKS LIKE PROPINQUITY: USING MACHINE LEARNING TO ESTIMATE THE EFFECTS OF SPATIAL PROXIMITY IN THE MAJOR LEAGUE BASEBALL DRAFT

Majid Ahmadi
Nathan Durst
Jeff Lachman
John A. List
Mason List
Noah List
Atom T. Vayalinkal

Working Paper 30786
http://www.nber.org/papers/w30786

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
December 2022

"Nothing Propinks Like Propinquity" is drawn from Ian Fleming's 1956 James Bond novel, Diamonds Are Forever. Thanks to Charles Shi for excellent research assistance–he and his remarkable summer team lent deep research assistance. Ahmadi, Vayalinkal, and the three Lists were creating a draft model for the Chicago White Sox (in partnership with Durst, Lachman, and Jeremy Haber) when these data were gathered. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

At least one co-author has disclosed additional relationships of potential relevance for this research. Further information is available online at http://www.nber.org/papers/w30786.ack

Nothing Propinks Like Propinquity: Using Machine Learning to Estimate the Effects of Spatial Proximity in the Major League Baseball Draft

Majid Ahmadi, Nathan Durst, Jeff Lachman, John A. List, Mason List, Noah List, and Atom T. Vayalinkal

NBER Working Paper No. 30786
December 2022
JEL No. C93,D4,J30,J7

ABSTRACT

Recent models and empirical work on network formation emphasize the importance of propinquity in producing strong interpersonal connections. Yet, one might wonder how deep such insights run, as thus far empirical results rely on survey and lab-based evidence. In this study, we examine propinquity in a high-stakes setting of talent allocation: the Major League Baseball (MLB) Draft from 2000-2019 (30,000 players were drafted from a player pool of more than a million potential draftees). Our findings can be summarized in four parts. First, propinquity is alive and well in our setting, and spans even the latter years of our sample, when higher-level statistical exercises have become the norm rather than the exception. Second, the measured effect size is consequential, as MLB clubs pay a significant opportunity cost in terms of inferior talent acquired due to propinquity bias: for example, their draft picks are 38% less likely to ever play a MLB game relative to players drafted without propinquity bias. Third, those players who benefit from propinquity bias fare better both in terms of the timing of their draft picks and their initial financial contract, conditional on draft order. Finally, the effect is found to be the most pronounced in later rounds of the draft, where the Scouting Director has the greatest latitude.

Majid Ahmadi
University of Chicago
mahmadi@uchicago.edu

Nathan Durst
Chicago White Sox, Head Scout
and National Cross-Checker
n_durst@msn.com

Jeff Lachman
Chicago White Sox, Head Scout
and National Cross-Checker
jeffrey.lachman@gmail.com

John A. List
Department of Economics
University of Chicago
1126 East 59th
Chicago, IL 60637
and Australian National University
and also NBER
jlist@uchicago.edu

Mason List
University of Chicago
mlist@ucls.uchicago.edu

Noah List
Harvard University
nlist@college.harvard.edu

Atom T. Vayalinkal
University of Toronto
Max Gluskin House
150 St. George Street
Toronto, ON
M5R 2T7
Canada
atom.vayalinkal@mail.utoronto.ca

Introduction

Spatial considerations have played a central role in economics for centuries. Within regional and urban economics, spatial composition of firms represents a key driver to both Marshallian agglomeration externalities as well as Jacobs externalities (Marshall, 1890; Jacobs, 1970), and has been linked to innovation spillovers (Jaffe et al., 1993). A workhorse model within international trade to predict bilateral trade flows is the gravity model, which is based on the economic size and distance between two countries to mimic gravitational interaction, as described in Newton's Law of Gravity (Isard, 1954).

A similar approach has been used to model immigration flows (Ramos and Royuela, 2016). In 1929, Hotelling introduced his competitive model to the Industrial Organization literature highlighting the relationship between location and prices. His key argument held that firms do not exercise product differentiation, rather they compete via prices based on geographic location (Hotelling, 1929). Space has even played a role in how much pleasure a specific action is theorized to induce. In his felicific calculus, Bentham (1789) surmised that the amount of hedonic pleasure from consumption was a direct function of its remoteness, or "propinquity"[1].

Of course, space has played a central role across other social sciences as well. In sociometrics, the concept of space can be linked to Durkheim's (1893) work. And, while in the early literature space occupied a back seat behind the traditional determinants of interpersonal connections (such as age, socioeconomic status, gender, and race (Bell, 1981), propinquity has played an increasingly important role in network formation. Two people working on the same floor of a building, for example, have a higher propinquity than two workers from different

[1] "Propinquity" comes from social psychology and means physical proximity (Nahemow and Lawton, 1975).

floors, just as two people with similar political beliefs possess a higher similarity than those whose beliefs strongly differ. There are now dozens of studies that show the power of propinquity in determining partnerships and friendships. A conclusion from this literature is that few variables are more important than propinquity when it comes to understanding friendship networks (see, e.g., Reagans, 2011).

In this study, we move the research question from the examination of people connected to people to one of exploring the allocation of people to organizations. Since a central feature of economics is the efficient allocation of talent, an important empirical question relates to whether propinquity is a factor within the matching of employee and employer in labor markets. Our specific empirical setting is a high-stakes environment whereby we examine the Major League Baseball (MLB) Player Draft from 2000-2019. Specifically, we explore the draft picks across every MLB club of the nearly 30,000 players drafted (from a player pool of more than a million potential draftees).

Our key analysis of propinquity revolves around the spatial similarity of the players under consideration and the Director of Scouting for each Major League club. Importantly, over this time period given that the draft was between 40 and 50 rounds, we can explore if players drafted earlier (who receive much more scrutiny than those drafted in later rounds) have less propinquity bias than those drafted in later rounds, where the scouting director has more latitude to drive the decision. We also employ several counterfactual exercises including when Scouting Directors change clubs, when they change residential locations, and when two different Directors work for teams located in the same city. In this sense, we are also afforded the opportunity to examine if the spatial similarity of the players under consideration and the Major League club

itself is important.

We report several insights. First, propinquity is alive and well in our setting, and spans even the latter years of our sample, when entire data teams using higher-level statistical exercises have become the norm rather than the exception in Major League baseball. For example, in our base model, a player is 7.1% more likely to be drafted by a particular team if he lives 1000km closer to the scouting director, controlling for skill. And, the player is 4.9% more likely to be drafted by a particular team if he lives 1000km closer to the city where that team plays. Second, the measured effect size is important, as MLB clubs pay a real cost in terms of inferior talent acquired due to propinquity bias: for example, their draft picks appear in 25 fewer games relative to teams that do not exhibit propinquity bias. In addition, in a counterfactual exercise we explore learning effects by examining Scouting Directors who change teams. We find that a Director is 9.9% more likely to draft a player who lives 1000km closer to the director, if the scouting director is on his first team, whereas the director is 14.4% more likely to draft a player who lives 1000km closer to him if the scouting director is on his second team (as directors move to a second team, getting closer on average). This result suggests Directors fail to learn about the bias over time and for their new job.

A third insight is that those players who benefit from propinquity bias also receive financial benefits: conditional on their draft order, their initial contracts are superior to counterfactual draft picks (by 12%-25%). We view this result as novel to the literature in that it shows not only propinquity in choice but also in prices, conditional on the biased choice. Finally, the effect is found to be the most pronounced in later rounds of the draft (after round 15), where the Scouting Director has the greatest latitude. For instance, for rounds 16+, our results

imply that a player is 11.4% more likely to be drafted by a team if he lives 1000km closer to the city where the team plays, and 11.8% more likely to be drafted by the team if he lives 1000km closer to the scouting director, controlling for player quality.

The remainder of our study begins with a discussion of the propinquity literature. We then summarize our data, the empirical approach, and our results in Section 3. Section 4 concludes.

1. Brief Literature Overview

Similarity and propinquity are deeply embedded in the decision-making processes of both individuals and firms alike. A facet that we focus on here is the social aspect of propinquity. A number of papers have documented individuals' preferences for others who are socially similar to them. In the personal sphere, adult friendship has been demonstrated to be homogeneous along age, gender, religion (Bell, 1981; McPherson et al., 2001), ethnicity, and even behaviors and ability (Hartup and Stevens, 1997; Allan, 1998). In recent years, there is also growing evidence that social relationships can also favor political similarity (Huber and Malhorta, 2017). Moreover, the preference for social similarity is consistent across different demographic groups, from college students (Godley, 2008) to older teachers (Reagans, 2011). Such observations can be explained in economic models of friendships as being due to both endogenous, homophilic preferences, and exogenous, differing opportunity sets (Currarini et al., 2009).

The effects of social similarities extend beyond friendship formations. A natural field experiment using correspondence tests on help-wanted advertisements found that African-American names elicit the same decrease in callbacks as the unrecognizable foreign-sounding names with no clear ethnic association, suggesting preference for ethnic similarity in the labor market is a large factor in ethnicity discrimination in the labor market

(Jacquemet and Yannelis, 2012). Another paper found evidence that doctors refer to specialists of their same gender, contributing to some of the earnings gap among physicians (Zeltzer, 2017). Indeed, there are important consequences in the labor market due to social propinquity. This type of relationship has even been found to affect charitable giving, as List and Price (2009) find evidence of similarities in race and gender affecting giving rates (though the economically significant result in their data is discriminatory in nature).

We view propinquity operating in the spatial, or physical, sense. Since the propinquity effect was first proposed in Festinger et al. (1950), numerous papers have documented the importance of physical propinquity in friendship formation among neighbors (Nahemow and Lawton, 1975; Reagans, 2011), subjects in lab experiments (Shin et al., 2019), and classmates (Kubitschek and Hallinan, 1998). Additionally, such physical nearness amplifies social ties among religious congregation members (Bulten, 2002) and teachers and school staff in elementary schools (Spillane et al., 2017), affects a firm's social influence among consumers (Meyners and Becker, 2017), and even likelihood of marriage (Bossard, 1932). This mechanism is likely due to physical propinquity increasing exposures between individuals or firms. In labs, children have shown preferences for friendship formation with children whose pictures they had more exposure to (Ball and Cantor, 1974; Liberman and Shaw, 2019), and people who have repeated contact with gay or lesbian individuals display friendlier attitudes towards them (Barth et al., 2009; Anderssen, 2002).

While physical propinquity has been found in networks, we view the work on investment decisions as related to our main hypothesis as well. An example is the well-documented home-bias effect among US investors, both internationally in favor of the United States (French and Poterba, 1991), domestically in favor of geographically proximate equities (Coval and

Moskowitz, 1999), and even in online transactions (Hortaçsu et al., 2009). Many papers attribute the bias to information asymmetry and high information costs (Coval and Moskowitz, 2001; Ahearne et al., 2004) as a reason, possibly due to investors profiting more from knowing information others do not know (van Nieuwerburgh and Veldkamp, 2009). Moreover, retail funders investing in artists on crowdfunding platforms also demonstrate a physical propinquity bias (Agrawal et al., 2015).

In the spirit of this literature, Scouting Directors might profit (odds of winning) from scouting in geographically proximate locations that they are familiar with through repeated exposure. While our data will not allow us to pinpoint which of the various mechanisms are at work, we do have rich enough data to explore for the presence of propinquity effects.

To gain a visual perspective of our data, we plot the distribution of talent and teams/directors in Figure 1. The figure is created by identifying qualified players using a reliable external source to grade baseball players (PerfectGame), which includes data for hundreds of thousands of baseball players and provides rankings for each player.[2] The figure shows the ranking of the state for the density of drafted players as well as the number of scouting directors and MLB teams in each state. The raw data reveal that the percent of players drafted from a state is correlated to the presence of MLB teams and scouting directors. Although the figure is only a visual representation of propinquity/proximity bias, various analysis will be conducted in this paper to analyze the existence and extent of propinquity bias in such data.

[2] In the figure, "qualified" players are those who have a grade of 9 or higher (out of 10) since the average grade for drafted players is 9.2 across different years. The data are also aggregated to the state level and includes Washington DC, Puerto Rico, and a few Canadian provinces. States/provinces are then numerically ordered based on the percent of drafted players from that state.

2. The Data, The Empirical Approach, and The Results

In this section we first describe our data and the various sources used. We then discuss our empirical approach to estimating the basic propinquity model and summarize several counterfactual exercises. Our empirical approaches are standard in the literature, but it should be noted that we do not control the assignment mechanism of our key treatment variables. As such, making strong causal statements requires untenable assumptions (see Harrison and List, 2004). We therefore refrain from structural language throughout.

2.1 The Data

Our data cover all players who have been drafted into professional baseball from 2000-2019. In total, this provides information on 27,679 players over these 20 years. We also collect performance data for more than 300,000 'draftable' players from sources such as Perfect Game, Game Changer, Prep Baseball Report (PBR), and NCAA data.

These data are compiled in a Machine Learning (ML) model to predict expected performance of players based on their high school and college profiles. We built this model for the Chicago White Sox to aid in their 2020 (and beyond) draft choices. The intuition behind the ML model is to divide the database into two sets of training and test. We build our model parameters based on the training dataset, and then evaluate the accuracy of our model on our test database. Therefore, by training Decision Trees for all measures, the remaining part of the database (test) is used for testing the accuracy of the model and making adjustments to the model accordingly. The developed model predicts various outcomes including expected draft round, probability of making it to the majors, and even Wins Above Replacement (WAR) score for each player. This model enables us to estimate a player's expected performance (player's quality).

Our last necessary data ingredient to test for propinquity bias is to determine physical distances between players from both high school and college to each Scouting Director and MLB team. As two of the co-authors were employed by an MLB club during the production of this research, we were able to use proprietary data to measure distances. This unique database provides the opportunity to conduct the proposed analysis of propinquity.

2.2 Empirical Approach and Results

At the most basic level, we examine our data by using a simple logistic regression model conditioning on factors that may influence whether a player is drafted:

$$\text{Drafted} = f\left(\alpha + \beta' X\right) \quad (1)$$

where Drafted equals 1 if a player was drafted, 0 otherwise; $f(\bullet) = 1/(1 + e^{-m})$ is the standard logit function; and X includes player-specific variables that may affect the propensity to be drafted as well as distance measures. Variables in X include various combinations of first year WAR (if players ultimately make it to the majors), first 7 seasons WAR, performance metrics in majors, and player's quality from our ML model. Moreover, two types of distances (player's distance to the team, and player's distance to the scouting director) were calculated.

The intuition behind this analysis is that we are conditioning on observable factors that affect perceived quality (both data available at the draft and thereafter, though if we just include results from our ML model the qualitative estimates remain the same) and are therefore testing for propinquity bias conditional on these controls. The resulting estimates provide a logs odd ratio that can be interpreted as the likelihood of being drafted based on your spatial location.

After estimating this base model, we then use a similar regression approach to explore various counterfactuals. These various counterfactual analyses are performed to check the robustness of findings and to provide different approaches to explore propinquity. For instance,

the propinquity effect can be measured when two MLB clubs share locations but their Directors of Scouting live in different cities, or when Scouting Directors change teams but they remain in the same location (we explored when Scouting Directors changed locations themselves but there was not a large enough sample to draw any conclusions.

2.3 Base Model: All-Director Analysis Across Rounds

This section starts with testing whether propinquity exists in the drafting process with inclusion of all directors throughout the entire draft. Table 1 provides summary estimates from equation (1). Summarized results include coefficients in the logistic regression that are marginal effects (one can take natural log(odds ratio)/1000 for inference).

Empirical results suggest that greater distance (both to the team and to the director) reduces the probability of being drafted under all sets of controls. Interpreting our results, for each model, we report the odds ratio and the relevant p-values. In context, this means that, for the first model, a player is 7.1% more likely to be drafted by a particular team if he lives 1000km closer to the scouting director, controlling for skill level, and 4.9% more likely to be drafted by a particular team if he lives 1000km closer to the city where that team plays.

Table 1 provides interesting initial insights, but the MLB draft experts understand that there are important asymmetries across rounds of the MLB draft. For example, there is a heterogeneity between players who are drafted in various rounds. In other words, a 1st round draft pick cannot be compared with a 30th round draft pick because of the heterogeneity in amount of exposure and attention each player receives: while players considered to be prospects for early rounds receive scrutiny from dozens of scouts for each team, those drafted later receive much less scrutiny. Therefore, Scouting Directors have much greater latitude in terms of drafting according to their

own preferences in later rounds of the MLB draft. To test if such latitude matters, we again estimate equation (1) but allow for heterogeneity across early, mid, and later rounds. Table 2 provides those estimates (Table 2A controls for first year WAR and 2B controls for first year ML model player prediction).

Again, we report odds ratio and p-value for distance in each model specification in Table 2. For rounds 16+, our results imply that a player is 11.4% more likely to be drafted by a team if he lives 1000km closer to the city where the team plays, and 11.8% more likely to be drafted by the team if he lives 1000km closer to the scouting director, controlling for first year war. Also, we report odds ratio and p-value for distance in each model specification. For rounds 16+, our results imply that a player is 6.5% more likely to be drafted by a team if he lives 1000km closer to the city where the team plays, and 4.0% more likely to be drafted by the team if he lives 1000km closer to the scouting director when using the ML Model Prediction for quality of skill.

Here, we find that the distance to team is a much stronger predictor of draft success in later rounds (16+) as opposed to early rounds (1-5), when controlling for player's quality (using scores from ML model). Accordingly, distance to director matters throughout, but is also a stronger predictor in later rounds. We should note that any round split one can imagine yields similar results qualitatively. For example, with the draft now limited to 20 rounds, we find considerably more propinquity in rounds 11-20 versus 1-10.

2.4 Counterfactual Analysis: Directors Changing Teams

A first robustness test we consider is to explore directors who have changed teams. This analysis creates the opportunity to test whether drafting patterns change when directors' MLB club changes location. In addition, this analysis shows whether there are level differences

between these directors and others, in terms of the average distance of players chosen.

To conduct the analysis, we should mention that of the 124 total MLB scouting directors from 2000-2019, 13 have worked for more than one team and all of those worked for exactly two teams. Of those 13, 7 moved to a team closer to their hometown, 5 moved further away, and 1 moved a few miles away. On average, MLB directors' hometowns are 917 km from their team city, however two-team directors are different. On their first team, their average distance is 1375 km while on their second team, their average distance is 1045 km. Thus, directors who have been employed by two-teams are characterized by having their second team being located closer to their home residence.

Then, the question becomes how distance to drafted players changes after their move. Interestingly, for the second team, the average player distance to a team is 15% lower (1612 km vs 1893 km) and average player distance to a director's hometown is 13% lower (1600 km vs 1844 km). In aggregate, this finding reveals that directors tend to draft players from nearby areas when they move closer to their hometowns.

We can move from anecdotes to a more formal data analysis by estimating equation (1) to explore how drafting decisions change due the scouting director move. Table 3 summarizes our empirical results.

Similar to our previous tables, we report the odds ratio *(exp(1000*coefficient))* and the p-value for each specification in Table 3. In this case, we also adjust for the cluster effect of directors in each regression. Interpreting the first model, controlling for player quality with first year WAR, a director is 9.9% more likely to draft a player who lives 1000km closer to the director, if the scouting director is on his first team, whereas the director is 14.4% more likely to draft a player who lives 1000km closer to him, if the scouting director is on his second team (as

directors move to a second team, getting closer on average). This supports our anecdotal finding that propinquity bias increased as directors moved to a second team. This suggests that player distance is a stronger predictor of draft probability on the second team, under all sets of controls.

2.5 Counterfactual Analysis: Cities with Two Teams

A second type of counterfactual analysis that we consider relates to multiple teams located in the same city. This permits us to focus on an empirical estimate that equalizes MLB club location but varies scouting director location. In other words, in these markets, two teams have access to the same pool of local players, but the scouting directors live in different cities. This analysis allows us to compare the strategies of these teams in how they select players.

To operationalize this idea, we consider 5 pairs of teams that are located in the same region: Chicago, New York, Los Angeles, California Bay Area, Baltimore/Washington DC. It is worth noting that Washington DC did not have a team prior to 2005 and players drafted more recently may not have reached MLB yet, so the original database was filtered for years 2005-2015. We then calculate the average draft distance for these teams for the second half of the draft (rounds 25+). These rounds have the highest amount of propinquity bias based on the findings from Table 2. Table 4 provides summary results.

As shown in Table 4, on average one of the teams (in each region) tends to draft players who live closer to the team, sometimes by more than several hundred kilometers. This can be considered as a treatment for drafting players from nearby areas (or propinquity treatment). Thus, in all 5 markets, the team on the right takes more nearby players than the team on the left.

2.6 Financial Gains: Signing Bonuses

Given that we have shown propinquity bias helps nearby players rise up the draft boards, one might ask whether there is a similar bias in terms of the financial package/signing bonuses provided to nearby players. To examine the financial dimension of the propinquity bias, we collected data on signing bonuses for all players drafted between 2000-2019 (this is the primary means of differing payments for minor league players since annual pay is homogeneous). Those players who did not have signing bonus data were excluded from our analysis, which still provides a sample size of over 5000 players. Then, the quality or value of a player to his team was assessed by using the ML player skill level model. By simple examination of the distribution of signing bonuses, we find that the data have a skewed distribution, with a mean of roughly $500,000 and a standard deviation above $800,000.

As such, we regress a logarithmic transformation of the signing bonuses via an OLS model (as well as a few other models) on different measures of distance between a player and director, allowing us to estimate equation (2):

$$\log(s) = f(\mathrm{ML}, d, \mathbb{I}\{\text{same region}\}) \tag{2}$$

Here, *s* is the signing bonus, ML is the machine learning model score, *d* is the distance to director and finally an indicator variable on whether the player and the director live in the same region (based on the Little League World Series map). Empirical results reveal that players living in the same region as the director are paid 12.76% higher signing bonuses (p-value .0410), controlling for distance and player quality using our ML model.

Then, we examine a more granular measure of the distance using a dummy variable if the player and director live in same state. Interestingly, empirical results of the second model reveal that players living in the same state as the director are paid 23.85% higher signing bonuses (p-value .0134), after controlling for distance and player quality.

Both models, however, ignore the influence of the draft round in determining signing bonuses; thus, a Two-Stage Linear Square (2SLS) model is utilized. In this model, we first estimate the round in which a player is predicted to be drafted using the ML model score, with a dummy variable indicating whether the player is being drafted out of high school or college, and the distance from player to director (in kilometers). In this first stage of the model, we find that distance is nearly statistically significant (p-value .0669) with a modest effect size (e.g., if a player is 1000 km closer, he is expected to be drafted 0.15 rounds earlier, controlling for ML score and HS/NCAA).

We then use this predicted round and the dummy variable (indicating if player and director live in the same state) to estimate the signing bonus. In this model, the state indicator is significant (p-value .0287) with the interpretation that players living in the same state as the director are paid 19.13% higher signing bonuses, controlling for predicted draft round. Summary results are contained in Table 5.

Overall, in each of these models, geographical and cultural proximity are statistically significant determinants of the player signing bonus, controlling for quality of players' talent, as shown in the above table. This shows that there is a positive financial reward for nearby players.

2.7 Cost of Propinquity

Beyond estimating the level and extent of propinquity in terms of draft picks and signing bonuses, we can estimate the cost of propinquity bias to the major league clubs. To provide such estimates, we perform a counterfactual analysis to compare the outcome for players who live geographically near the team/director, compared to other players after controlling for player quality. For this analysis, we create a new variable to express distance to players, which is called

"playerpct". This variable is defined as:

$$Playerpct_{i,t} = \frac{distance\ from\ the\ player\ to\ the\ team}{\text{Average distance from the team to ALL players drafted}}\ \forall player\ i\ \&\ team\ t$$

This variable is calculated for each player (i) and team (team), and then a subset of the data on rounds 11+ (where the propinquity is at the highest as shown earlier) and draft years 2000-2016 was selected. Years (2017-2019) were excluded due to recency, since many of these players are still in minor leagues.

A treatment indicator was defined as a dummy variable being equal to 1 if "playerpct" is less than 0.1, and 0 otherwise. This 10% distance is between 150 to 250 km for most teams, though it varies by team and is largest for teams in the northwest, who are furthest from the largest cluster of talent in the southeast. Subsequently, a logistic regression model was applied to predict the binary outcome of a draft pick reaching the MLB, as a function of the player's quality (ML model score), a dummy for high school vs. college, team fixed effect, playerpct, and treatment as written below:

$$\log(MLB) = f(ML\ Model\ Score, HS\ dummy, team\ fixed\ effect, playerpct, treatment) \quad (3)$$

The treatment dummy variable which indicates if a player was within 10% of the average distance from the team to all available players, is the variable of interest when estimating equation (3). The goal is to identify the cost that teams bear for employing a strategy of severe propinquity bias, where scouting directors are targeting players with the shortest proximity.

When we estimate this model, which includes 5708 observations, we find that the treatment dummy variable is significant with a p-value of 0.018, yielding an odds ratio that a player is 62% as likely to reach MLB if he was within 10% of the average distance from that team, controlling for skill. In other words, such a player is 38% less likely to ever play an MLB game. As such, this model reveals a consequential cost of propinquity bias (complete results are

available upon request).

From 2000-2016, 109 players drafted in rounds 11+ [from within 10% of the average available distance from the teams who drafted them] reached the MLB. If 61.6% more of them were successful, that would yield an additional 68 players for these years, or 4.25 players per draft year. Given that most draft picks will not play any games at the Major League, the expected differences exhibited in our results are quite large. The strength of this model should also be emphasized. If we consider the fitted values of the logistic regression as predictions, taking a prediction greater than 0.50 to be an estimate of reaching the MLB, and a fitted value below 0.50 to be an estimate of not reaching the MLB, then the predictions are 89.7% accurate.

2.8 Cluster Analysis

To enhance our understanding of the heterogeneity between various strategies that scouting directors and teams have employed across different years, in this section of the paper a clustering analysis is conducted. In general, cluster analysis is used to identify similar groups within a population by minimizing the distance/difference between members of each identified group. Here a wide range of clustering analysis, including hierarchical and K-means clustering, are employed. The hierarchical clustering returned more interpretable and meaningful results which are explained below. To perform the hierarchical clustering, we use Ward's method, which works based on minimizing the variation within clusters. In addition, distances between group members are measured based on the Euclidean distances. Overall, the hierarchical clustering reveals having 6 unique groups of players based on their performance and draft rounds.

Table 6 summarizes high level statistics of these clusters based on the performance metrics of players in each cluster, WAR score, distance to the team and director. The table begins to add

clarity on the different strategies for drafting players. In particular, focusing on cluster 4, which has the lowest WAR stats and the highest average draft round, the average draft round for this cluster is 31, which falls into the 2nd half of draft rounds. Interestingly, for players in this cluster the average distance to the director and team are relatively small, which confirms the propinquity bias studied in the previous sections. In other words, cluster 4 shows one of the existing drafting strategies that involves taking players from nearby in late rounds, yielding unimpressive career outcomes for those players.

3. Epilogue

Prior to the rise of higher-level statistics, analytics, and computing power, baseball teams were entirely reliant on scouting to evaluate talent. Today, while a few teams spend hundreds of millions of dollars on free agents, most teams are dependent on the amateur draft to acquire players with the potential to be stars. Unlike in other sports, however, baseball draft picks often need several years of development in the minor leagues before they are ready to compete at the highest level. Because it can take a while for players to reach their potential, scouting for the amateur draft is not just about evaluating how players are performing currently, but about projecting what they will be able to perform in their prime.

History has shown that these projections are far from perfect, with factors such as height, race, and climate influencing scouts' estimates of how well players will perform in the future. With so many variables creating bias in draft decisions, we hypothesize that propinquity (geographical proximity) influences the evaluations that scouts develop on players. Different types of propinquity or locational proximity have been studied by researchers in different fields but to our knowledge the exploration of talent allocation to organizations in a high stakes setting is entirely fertile ground.

This paper focuses on the MLB draft for the first 20 years of this century. With an ability to control for several features of ability, we use a regression approach to examine the effects of propinquity in this setting. We report interesting results that show the reach of propinquity might be beyond what even the most optimistic propinquity theorists might have suspected. We hope that future studies explore the generalizability of our insights, focusing not only on allocation but prices. Recent models of generalizability (List, 2022) rely on the notion that differences in preferences, beliefs, and constraints across settings critically influence the portability of empirical insights. In this light, replications need to be completed to understand if the allocation and price result can be applied to other populations of people and situations. Given the high stakes of the MLB player draft, we suspect that our results will provide a good mapping, if not lower bound, to other settings.

References

Agrawal, Ajay, Christian Catalini, and Avi Goldfarb. "Crowdfunding: Geography, Social Networks, and the Timing of Investment Decisions." *Journal of Economics & Management Strategy* 24, no. 2 (June 2015): 253–74. https://doi.org/10.1111/jems.12093.

Ahearne, Alan G, William L Griever, and Francis E Warnock. "Information Costs and Home Bias: An Analysis of US Holdings of Foreign Equities." *Journal of International Economics* 62, no. 2 (March 1, 2004): 313–36. https://doi.org/10.1016/S0022-1996(03)00015-1.

Allan, Graham. "Friendship, Sociology and Social Structure." *Journal of Social and Personal Relationships* 15, no. 5 (October 1998): 685–702. https://doi.org/10.1177/0265407598155007.

Anderssen, Norman. "Does Contact with Lesbians and Gays Lead to Friendlier Attitudes? A Two Year Longitudinal Study." *Journal of Community & Applied Social Psychology* 12, no. 2 (March 2002): 124–36. https://doi.org/10.1002/casp.665.

Ball, Portia M., and Gordon N. Cantor. "White Boys' Ratings of Pictures of Whites and Blacks as Related to Amount of Familiarization." *Perceptual and Motor Skills* 39, no. 2 (December 1974): 883–90. https://doi.org/10.2466/pms.1974.39.2.883.

Barth, Jay, L. Marvin Overby, and Scott H. Huffmon. "Community Context, Personal Contact, and Support for an Anti—Gay Rights Referendum." *Political Research Quarterly* 62, no. 2 (June 2009): 355–65. https://doi.org/10.1177/1065912908317033.

Bell, Robert R. "Friendships of Women and of Men." *Psychology of Women Quarterly* 5, no. 3 (March 1981): 402–17. https://doi.org/10.1111/j.1471-6402.1981.tb00582.x.

Bentham, Jeremy. *An Introduction to the Principles of Morals and Legislation*. Oxford: Oxford University Press, 1789.

Bulten, Tom. "Community and Propinquity of Church Members." *Christian Scholar's Review* 31, no. 4 (Summer 2002): 359–75.

Coval, Joshua D., and Tobias J. Moskowitz. "Home Bias at Home: Local Equity Preference in Domestic Portfolios." *The Journal of Finance* 54, no. 6 (December 1999): 2045–73. https://doi.org/10.1111/0022-1082.00181.

———. "The Geography of Investment: Informed Trading and Asset Prices." *Journal of Political Economy* 109, no. 4 (2001): 811–41. https://doi.org/10.1086/322088.

Currarini et al "An Economic Model of Friendship: Homophily, Minorities, and Segregation." *Econometrica* 77, no. 4 (2009): 1003–45. https://doi.org/10.3982/ECTA7528.

Durkheim, Émile. *The Division of Labor in Society*. New York: Free Press, 1893.

Festinger, Leon, Stanley Schachter, and Kurt Back. *Social Pressures in Informal Groups: A Study on Human Factors in Housing*. Stanford, Calif: Stanford Univ. Press, 1950.

French, Kenneth R., and James M. Poterba. "Investor Diversification and International Equity Markets." *The American Economic Review* 81, no. 2 (1991): 222–26. https://www.jstor.org/stable/2006858.

Godley, Jenny. "Preference or Propinquity? The Relative Contribution of Selection and Opportunity to Friendship Homophily in College." *Connections* 1, no. 1 (2008): 65–80.

Harrison, Glenn W., and John A. List. "Field Experiments." *Journal of Economic Literature* 42, no. 4 (December 2004): 1009–55. https://doi.org/10.1257/0022051043004577.

Hartup, Willard W., and Nan Stevens. "Friendships and Adaptation in the Life Course." *Psychological Bulletin* 121, no. 3 (1997): 355–70. https://doi.org/10.1037/0033-2909.121.3.355.

Hortaçsu, Ali, F. Asís Martínez-Jerez, and Jason Douglas. "The Geography of Trade in Online Transactions: Evidence from EBay and MercadoLibre." *American Economic Journal: Microeconomics* 1, no. 1 (February 2009): 53–74. https://doi.org/10.1257/mic.1.1.53.

Hotelling, Harold. "Stability in Competition." *The Economic Journal* 39, no. 153 (1929): 41–57. https://doi.org/10.2307/2224214.

Huber, Gregory A., and Neil Malhotra. "Political Homophily in Social Relationships: Evidence from Online Dating Behavior." *The Journal of Politics* 79, no. 1 (January 2017): 269–83. https://doi.org/10.1086/687533.

Isard, Walter. "Location Theory and Trade Theory: Short-Run Analysis." *The Quarterly Journal of Economics* 68, no. 2 (May 1954): 305. https://doi.org/10.2307/1884452.

Jacobs, Jane. *The Economy of Cities*. New York: Vintage Books, 1970. Jacquemet, Nicolas, and Constantine Yannelis. "Indiscriminate Discrimination: A Correspondence Test for Ethnic Homophily in the Chicago Labor Market." *Labour Economics* 19, no. 6 (December 2012): 824–32. https://doi.org/10.1016/j.labeco.2012.08.004.

Jaffe, Adam B., Manuel Trajtenberg, and Rebecca Henderson. "Geographic Localization of Knowledge Spillovers as Evidenced by Patent Citations." *The Quarterly Journal of Economics* 108, no. 3 (1993): 577–98. https://doi.org/10.2307/2118401.

Kubitschek, Warren N., and Maureen T. Hallinan. "Tracking and Students' Friendships." *Social Psychology Quarterly* 61, no. 1 (March 1998): 1. https://doi.org/10.2307/2787054.

Liberman, Zoe, and Alex Shaw. "Children Use Similarity, Propinquity, and Loyalty to Predict Which People Are Friends." *Journal of Experimental Child Psychology* 184 (August 2019): 1–17. https://doi.org/10.1016/j.jecp.2019.03.002.

List, John A, "Non est Disputandum de Generalizability? A Glimpse into The External Validity

Trial," Technical Report, National Bureau of Economic Research (2020).

List, John A., and Michael K. Price. "The Role of Social Connections in Charitable Fundraising: Evidence from a Natural Field Experiment." *Journal of Economic Behavior & Organization* 69, no. 2 (February 2009): 160–69. https://doi.org/10.1016/j.jebo.2007.08.011.

McPherson, Miller, Lynn Smith-Lovin, and James M Cook. "Birds of a Feather: Homophily in Social Networks." *Annual Review of Sociology* 27, no. 1 (August 2001): 415–44. https://doi.org/10.1146/annurev.soc.27.1.415.

Meyners, Jannik, Christian Barrot, Jan U. Becker, and Jacob Goldenberg. "The Role of Mere Closeness: How Geographic Proximity Affects Social Influence." *Journal of Marketing* 81, no. 5 (September 2017): 49–66. https://doi.org/10.1509/jm.16.0057.

Nahemow, Lucille, and M. Powell Lawton. "Similarity and Propinquity in Friendship Formation." *Journal of Personality and Social Psychology* 32, no. 2 (1975): 205–13. https://doi.org/10.1037/0022-3514.32.2.205.

Ramos, Raul, and Vicente Royuela. "'Graduate Migration in Spain: The Impact of the Great Recession on a Low Mobility Country.'" AQR Working Papers. University of Barcelona, Regional Quantitative Analysis Group, April 2016. https://ideas.repec.org/p/aqr/wpaper/201608.html.

Reagans, Ray. "Close Encounters: Analyzing How Social Similarity and Propinquity Contribute to Strong Network Connections." *Organization Science* 22, no. 4 (August 2011): 835–49. https://doi.org/10.1287/orsc.1100.0587.

"Residential Propinquity as a Factor in Marriage Selection." *American Journal of Sociology* 38, no. 2 (September 1932): 219–24. https://doi.org/10.1086/216031.

Shin, Ji-eun, Eunkook M. Suh, Norman P. Li, Kangyong Eo, Sang Chul Chong, and Ming-Hong Tsai. "Darling, Get Closer to Me: Spatial Proximity Amplifies Interpersonal Liking." *Personality and Social Psychology Bulletin* 45, no. 2 (February 2019): 300–309. https://doi.org/10.1177/0146167218784903.

Spillane, James P., Matthew Shirrell, and Tracy M. Sweet. "The Elephant in the Schoolhouse: The Role of Propinquity in School Staff Interactions about Teaching." *Sociology of Education* 90, no. 2 (April 2017): 149–71. https://doi.org/10.1177/0038040717696151.

Van Nieuwerburgh, Stijn, and Laura Veldkamp. "Information Immobility and the Home Bias Puzzle." *The Journal of Finance* 64, no. 3 (June 2009): 1187–1215. https://doi.org/10.1111/j.1540-6261.2009.01462.x.

Zeltzer, Dan. "Gender Homophily in Referral Networks: Consequences for the Medicare Physician Earnings Gap." *SSRN Electronic Journal*, 2017. https://doi.org/10.2139/ssrn.2921482.

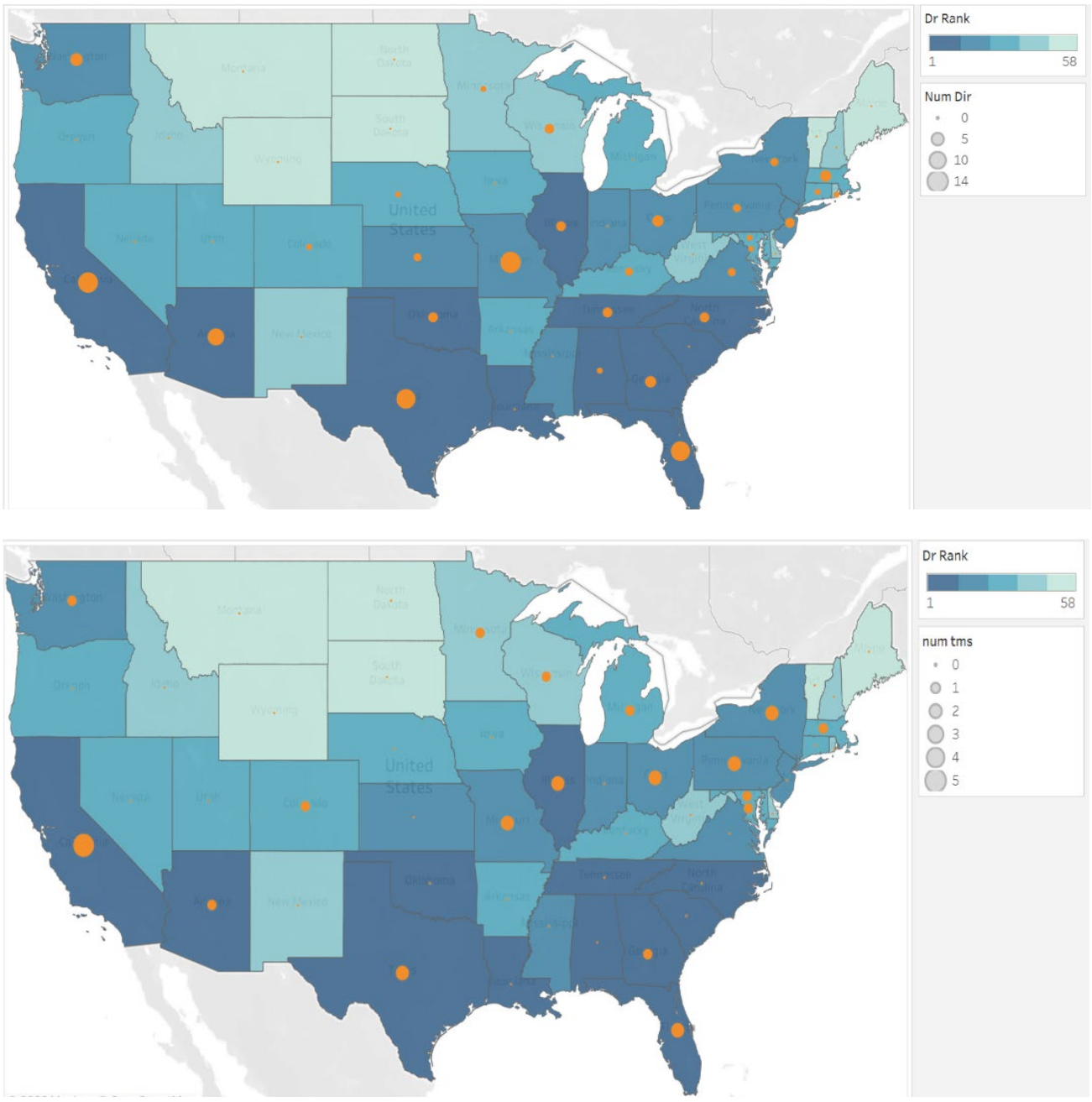

FIGURE 1: Drafted Players Spatial Association with Directors (top panel) and Teams (bottom panel)

Table 1: Propinquity Estimates for the Base Model

Logistic Regression: Importance of Player Distance on Draft Probability

Controls Used	Explanatory Variable	
	(1) Player-Team Distance	(2) Player-Director Distance
(1) First Year WAR	1.049 p=0.001*** N = 101,850 HLT p-value = 0.000	1.071 p=0.000*** N = 101,850 HLT p-value = 0.001
(2) First 7 Years WAR	1.049 p=0.001*** N = 101,850 HLT p-value = 0.000	1.071 p=0.000*** N = 101,850 HLT p-value = 0.001
(3) Performance Metrics (g, ab, hr, etc.)	1.049 p=0.038*** N = 43,350 HLT p-value = 0.002	1.078 p=0.002*** N = 43,350 HLT p-value = 0.322
(4) ML Model Prediction	1.04 p=0.000*** N = 335,970 HLT p-value = 0.000	1.038 p=0.000*** N = 335,970 HLT p-value = 0.000

Notes: This table presents regression estimates of equation (1) with various controls included and the primary focus on player-team distance (column 1) and player-director distance (column 2). Hosmer-Lemeshow Test (HLT) p-values are used as a measure of goodness of fit.

Table 2a: Propinquity Heterogeneity Across Rounds

Logistic Regression: Importance of Player Distance on Draft Probability, by Draft Round

Specification	Explanatory Variable (incl. First Year WAR Control)	
	(1) Player-Team Distance	(2) Player-Director Distance
(1) Draft Rounds 1-5	1.006 p=0.792 N = 39,450 HLT p-value = 0.036	1.044 p=0.084* N = 39,450 HLT p-value = 0.061
(2) Draft Rounds 6-15	1.034 p=0.256 N = 26,670 HLT p-value = 0.419	1.054 p=0.083* N = 26,670 HLT p-value = 0.238
(3) Draft Rounds 16+	1.114 p=0.000*** N = 35,730 HLT p-value = 0.004	1.118 p=0.000*** N = 35,730 HLT p-value = 0.013

Notes: This table presents regression estimates of equation (1), controlling for first year WAR, allowing heterogeneity across early, mid, and late rounds of the draft, with a primary focus on player-team distance (column 1) and player-director distance (column 2).

Table 2b: Propinquity Heterogeneity Across Rounds

Logistic Regression: Importance of Player Distance on Draft Probability, by Draft Round

Specification	Explanatory Variable (incl. ML Model Prediction Control)	
	(1) Player-Team Distance	(2) Player-Director Distance
(1) Draft Rounds 1-5	1.002 p=0.900 N = 57,930 HLT p-value = 0.804	1.042 p=0.045** N = 57,930 HLT p-value = 0.721
(2) Draft Rounds 6-15	1.012 p=0.467 N = 81,840 HLT p-value = 0.014	1.031 p=0.079* N = 81,840 HLT p-value = 0.025
(3) Draft Rounds 16+	1.065 p=0.000*** N = 196,200 HLT p-value = 0.000	1.04 p=0.000*** N = 188,880 HLT p-value = 0.000

Notes: This table presents regression estimates of equation (1), controlling for ML prediction, allowing heterogeneity across early, mid, and late rounds of the draft, with a primary focus on player-team distance (column 1) and player-director distance (column 2).

Table 3: Propinquity Measures: Directors Changing Teams

Logistic Regression: Importance of Player Distance on Draft Probability (Two-Team Directors)

Controls Used	**Explanatory Variable**			
	(1) Player-Team Distance		(2) Player-Director Distance	
	Team 1	Team 2	Team 1	Team 2
(1) First Year WAR	1.026 p=0.692 N = 9,664 HLT p-value = 0.38	1.14 p=0.022** N = 10,278 HLT p-value = 0.272	1.099 p=0.013** N = 9,664 HLT p-value = 0.904	1.144 p=0.001*** N = 10,278 HLT p-value = 0.736
(2) First 7 Years WAR	1.025 p=0.695 N = 9,664 HLT p-value = 0.091	1.141 p=0.021** N = 10,278 HLT p-value = 0.304	1.099 p=0.014** N = 9,664 HLT p-value = 0.23	1.144 p=0.001*** N = 10,278 HLT p-value = 0.703
(3) ML Model Prediction	1.05 p=0.446 N = 12,477 HLT p-value = 0.394	1.053 p=0.048** N = 39,165 HLT p-value = 0.014	1.033 p=0.585 N = 12,477 HLT p-value = 0.143	1.06 p=0.024** N = 39,165 HLT p-value = 0.045

Notes: This table presents regression estimates of equation (1) using data only for directors who changed teams, with various controls included and the primary focus on player-team distance (columns 1 and 2) and player-director distance (columns 3 and 4).

Table 4: Propinquity Measures: Cities with Two Teams

Region	Average Distance to Team (km)	
Chicago	Cubs	White Sox
	1474	1465
New York	Yankees	Mets
	1949	1947
Los Angeles Metro	Angels	Dodgers
	2264	2028
California Bay Area	Giants	Athletics
	2602	2375
Baltimore + Washington DC	Orioles	Nationals
	1809	1692

Notes: This table presents the average distance between team location and player.

Table 5: Propinquity in Terms of Signing Bonus

Signing Bonus Bias

Explanatory Variable	Signing Bonus Increase	p-value
(1) Same Director Region	12.76%	0.041^{**} N = 5049 $R^2 = 0.35$
(2) Same Director State	23.85%	0.0134** N = 5049 $R^2 = 0.35$
(3) 2SLS Director State	19.13%	0.0287** N = 5041 $R^2 = 0.37$

Notes: This table presents regression estimates of equation (2) with controls for player quality (rows 1 and 2) and second-stage estimates from the 2SLS model (row 3).

Table 6: A Cluster Approach to Propinquity

Cluster	Round (Mean)	Player-Director Distance (Mean)	Player-Team Distance (Mean)	WAR (Mean)	War (Var)	G (Mean)	AB (Mean)	HR (Mean)
1	25.4	1405.1	1775	0.4	3.7	50.1	93.2	1.8
2	9.1	933.8	927	1.5	26.7	80.4	222.2	8.9
3	11.6	2127.1	2075	1.8	17.3	102.2	260.4	9.6
4	30.7	772.1	658	0.3	2.8	38.1	80.3	3.7
5	10.4	696.3	609	1.5	14.5	110.9	250.6	10
6	20	478.3	558	1.5	21.1	75.8	224.3	10.8

Notes: This table presents the result of clustering analysis and identifying 6 different groups of players based on their performance metrics, WAR score, and distance to the team and director. The result confirms the existence of propinquity bias by identifying a unique group of drafted players that have worst performance and relatively low distance to the team & director (cluster 4).

Cluster	G.1 (Mean)	W (Mean)	L (Mean)	ERA (Mean)	SV (Mean)	BA (Mean)	OPS (Mean)	WHIP (Mean)
1	49.7	4.6	4.5	5.4	0.8	0.2	0.5	1.4
2	46.5	6.2	6.6	6.6	1	0.2	0.5	1.7
3	80.1	9	8.6	4.9	2.4	0.2	0.5	1.5
4	42.6	3.5	3.2	6.9	2.6	0.1	0.4	1.7
5	56.6	7.4	8.4	5.1	1	0.2	0.5	1.5
6	64.7	8	7.5	3.9	2.1	0.2	0.6	1.2