Quantitative Methods for Architecture Research: Lessons from the Social Sciences

Introduction

Architecture research can benefit from quantitative research methods commonly used in the social sciences. While there are limits to the validity of methods imported from other disciplines, architecture research, like that conducted in the social sciences, often deals with a messy amalgam of rather easily quantifiable variables (energy efficiency or levels of voter participation, for example), as well as somewhat more complex constructs involving what economist and psychologist Herbert Simon categorized as the "bounded rationality" of human subjects (Simon, 1991). These constructs ("confidence" or "happiness", for example), involve subjectivity and the occasional but nevertheless reliable irrationality of human behavior and sentiment, and they may therefore prove challenging—but not impossible—to convert into measurable variables. Indeed, there is a rich history of quantitative research in the social sciences, particularly in sociology and management, which aims to do precisely that (Stinchcombe, 1959; Lawrence & Lorsch, 1967; Cohen, March, & Olsen, 1972; Hannan & Freeman, 1977; Carroll & Delacroix, 1982; Romanelli, 1989; Amburgey & Dacin, 1994; Thornton & Ocasio, 1999; Glynn & Abzug, 2002; Dunn & Jones, 2010). In such scholarship, researchers begin with a broad research question involving a causal relationship of constructs ("*How does confidence affect happiness?*"), use extant literature to build hypotheses around that question ("*Increased levels of confidence will be associated with increased levels of happiness*"), and then test those hypotheses by developing measures that move from theoretical constructs to quantitative measures, through a process called operationalization (Trochim & Donnelly, 2008; Bono & McNamara, 2011).

One might, for example, operationalize "confidence" or "happiness" by conducting surveys or by using other variables such as heart rate, posture, or, perhaps, ear-to-ear grins

per minute (an admittedly dubious operationalization) as proxies for the constructs involved in the study. Such measures are inevitably imperfect, as they may embody and expose the limitations and cognitive biases of the research methods and the researcher him or herself. But they may also, however imperfectly, help us better understand some aspect of the world around us, and they may prove useful for other researchers who aim to correct, critique, or build upon the measures we have used.

In what follows, I will conduct an original, demonstration study of the relationship of health, wellbeing, ethnic diversity, and investment in the built environment, using data from various municipal departments of New York City, at the level of the 59 individual Community Districts (CDs) that comprise the city. While both the data and the statistical tests are real, the purpose of this study is to demonstrate a quantitative approach to research on the built environment, rather than to contribute to the literature around the research questions asked here. The social science techniques I employ here have seldom~~, if ever,~~ been used in the architecture literature, yet I believe they might prove useful in undertaking more rigorous scholarship, and in fostering greater exchange of findings and theoretical positions among scholars in the field~~.~~ This demonstration is intended to show how and why that might be so.

Research Question And Hypotheses

Con formato: Fuente: Sin Cursiva

Most scholarship begins with a research question, typically expressed in general terms. That is, the question does not allude to the actual population being studied or the specific data being used (*"Do 5-year-old children who exhibit signs of self-confidence smile more frequently than those that do not?"*), but instead describes the researcher's interest in the relationship of two or more constructs or theoretical concepts ("*How does confidence affect happiness?"*). A good research question should be relevant to current scholarship, and it should be complex enough that it engages a broader theoretical debate or question (Creswell, 2003). It should, however,

be simple enough that a relatively uninitiated reader can understand what is being asked. Perhaps most crucially, a research question should be *interesting* (Davis, 1971): it should aim to challenge or amend ~~or~~ previous scholarship, and not merely confirm the validity of others' findings (Miller and Salkind, 2002) .

The research questions to be addressed in this study are as follows: how does investment in the built environment affect the level of healthiness of those environments? Is public investment more likely to be associated with healthier environments than private investment? To this inquiry, I will add two parallel questions: controlling for the level of investment in the built environment, are communities with greater ethnic diversity healthier than those with lower levels of ethnic diversity? And do communities that experience a decline in environmental health over time have lower levels of social wellbeing? These research questions immediately require the researcher to define the constructs involved, namely, "investment in the built environment", "healthy environments", "ethnic diversity", and "social wellbeing". In a complete research article, one would ground these definitions in the literature, either to extend or critique the earlier approaches. For the purpose of this demonstration, and in the interest of brevity, I will simply define the constructs as follows:

Built Environment Investment (BEI): the degree to which resources are dedicated to constructing or maintaining the physical infrastructure in a community.

Con formato: Fuente: Sin Negrita

Environmental Health (EH): the degree to which the physical environment in a community contributes to human health and wellbeing.

Con formato: Fuente: Sin Negrita

Ethnic Diversity (ED): the degree to which a community is composed of people of different cultural, linguistic, ancestral, and/or national backgrounds.

Con formato: Fuente: Sin Negrita

Social Wellbeing (SW): the degree to which the members of a community have access to health care, education, and sufficient income to provide for material wellbeing.

Con formato: Fuente: Sin Negrita

These constructs *themselves* cannot be observed; they require operationalization in order to be transformed into measurable variables. Also, these definitions do not express *how* the terms are to be operationalized; there may be several equally valid methods to convert the constructs into variables, as will be illustrated later.

Research questions are most often expressed in general terms. ~~Although New York City is the setting for the current study,~~For the type of research presented in this demonstration, it should be the goal of the researcher to contribute a general principle about, for example, the relationship of investment and healthy environments, and not to describe a particular phenomenon in a particular place. Other modes of scholarship, especially qualitative case-study approaches, may dwell on the specifics of a given research setting, but even then will often make broader theoretical claims that extend beyond the specifics of the case at hand. In a full research article, one would be obliged to defend the generalizability of the case, and if the researcher were unable to do so, it would undermine considerably the potential contribution of the study (Bono & McNamara, 2011). For the purposes of this demonstration, I assume that the sheer size and diversity of New York City may provide a generalizable case from which to make broader observations and conclusions.

Hypotheses

Con formato: Fuente: Sin Negrita, Cursiva

~~:~~In a full research article, the author would use an extensive overview of the existing literature to argue for a series of hypotheses based on a presumed causal relationship among the constructs identified above (Sparrowe & Mayer, 2011). Because the objective here is to discuss

Con formato: Fuente: Sin Negrita, Cursiva

methodology, I will propose several hypotheses, but will not include a literature review or theoretical justification.

> ***Hypothesis 1a: (H1a):*** *Communities with higher levels of Built Environment Investment will have higher levels of Environmental Health.*
>
> ***Hypothesis 1b (H1b:)*** *The positive effect of Built Environment Investment on Environmental Health will increase as the proportion of investment derived from public work increases.*
>
> ***Hypothesis 2 (H2):*** *Communities with higher levels of Ethnic Diversity will have higher levels of Environmental Health.*
>
> ***Hypothesis 3 (H3):*** *Communities that have experienced a decline in Environmental Health over time will have lower levels of Social Wellbeing after the period of decline than those that experienced no such decline.*

When formulating hypotheses, researchers should write them in such a way that they can, in fact, be measured, and that they propose a correlation between constructs ("subjects with *x* will have increased *y*"). While it is possible for a researcher to formulate hypotheses that involve causation, ~~i~~In the type of research demonstrated here, I have taken care to predict a correlation ~~-~~without proposing a causal explanation in the formulation of the hypothesis itself ("subjects with *x* will have increased *y* because *x* stimulates production of *y*"). In ~~this type of researcha~~ study such as the one I demonstrate here , It is the role of the theoretical framework being developed in the article to provide a convincing explanation for causation; the hypotheses simply predict a correlation among variables.

Hypotheses should emerge from and be supported by the existing literature, from our own lived experience, and from a logical analysis of the question at hand. Yet they should also suggest something new or surprising about the phenomenon being studied (Davis, 1971; Sparrowe & Mayer, 2011). Unless it serves as a baseline for further hypotheses, a hypothesis that merely confirms existing theory or that predicts something obvious ("eating will result in reduced levels of hunger") is unlikely to be of

much value, and a study that consists solely of such hypotheses will seldom be published. The most important potential contribution of the article is often articulated in the hypotheses themselves.

Data And Variables

Empirical research requires data in order to test hypotheses. Choosing which data to use, and how much ~~of it~~ to collect, is a subjective part of the research process, but there are nevertheless a few useful rules of thumb that can be used when making decisions about data. First, researchers should be realistic about the resources at their disposal, and design a study that can reasonably be completed with those resources. In some cases, a researcher may have both the time and financial resources to collect an original data set, via fieldwork or archival research. Alternatively, existing data sets can be used if available, though the scientific contribution of such a study may be considered to be somewhat less significant than one using original data. Data selection should also be representative of he phenomenon of interest. If, for example, a researcher wished to conduct a survey on the possible correlation of automobile ownership and individuals' political opinions, he or she would, at the very least, want to ensure that data were collected only among those old enough to drive or own an automobile.

Data for this study have been collected from the New York City municipal government. I have used these data to generate the dependent, independent, and control variables, and to test the regression models. Regression, as I will explain in greater detail below, is a technique for estimating a relationship among variables, and a multiple regression model, such as the one employed here, includes several independent variables in a complex equation. (Lattin, Caroll, and Green, 2003). Regression models may describe a linear relationship among variables, in which an increase or decrease in a dependent variable is associated with an

Con formato: Fuente: Sin Cursiva

Comentado [A1]: Changed per editorial suggestion.

Comentado [A2]: Need to rephrase this sentence and add references here. Regression includes both linear and nonlinear forms, and normally multiple regression deals with linear models, but multiple nonlinear regression is also often mentioned.

Comentado [A3R2]: OK – changes made accordingly.

increase or decrease in an independent variable or variables. Nonlinear regression, meanwhile, describes a relationship between independent and dependent variables that cannot be described by a straight line, such as a U-shaped distribution or an exponential relationship among variables. While both types of regression are relevant to research in the social sciences (Agresti, 2018), in this demonstration I will work only with linear regression. Regression is a fundamental tool for quantitative analysis, and simple regression modelling can be done with the STATA application, as in this demonstration, or with other programs such as SPSS, R, or Excel, through available plug-in tools.

Dependent Variables:

Environmental Health (EH): ~~In a conventional research study, one would generally try to use a single operationalization across multiple hypotheses. Here, however,~~ I have used two different operationalizations in order to suggest the range of possible ways a researcher might reasonably approach the task of moving from construct to variable, and because it can be useful to operationalize a variable in different ways in order to test the robustness of one's findings. A second operationalization can either support or challenge a reseracher's results: should distinct operationalizations respond in a similar manner, we could allay some potential concerns over the validity of the measures being used. Contradictory results, however, may lead the researcher to question these measures.

In this demonstration, ~~-~~ I have used proxy variables in order to measure EH. Proxies can help researchers develop quantifiable variables by standing in for phenomena that may be difficult or impossible to quantify. The proxy variable itself is not necessarily of interest to the researcher. Instead, the proxy serves as a convenient (and plausible) substitute for something that resists measurement. In this demonstration, for example, I have used two proxy variables to estimate EH: The first is the variable *child_asthma*, which measures the number of

Con formato: Fuente: Sin Negrita

Comentado [A4]: This is a valuable part of what you're presenting—using quantifiable variables for something difficult to quantify-- but please explain this.

Comentado [A5R4]: OK, explanation added

Con formato: Sangría: Primera línea: 1,27 cm

childhood asthma hospitalizations (ages 5-14) per 10,000 inhabitants, per New York City Community District (CD) over the period 2012-2013. A second proxy is the variable *life_expect,* which measures the median life expectancy, in years, per CD. I use the first operationalization in tests of H1a and H1b, and the second in tests of H2. ~~One would generally avoid multiple operationalizations such as these in a single study; I do so here only to demonstrate the various options available to the researcher~~ It can be useful to operationalize a variable in different ways in order to test the robustness of one's findings; if a researcher were to develop.

Social Wellbeing (SW): I have used the American Human Development Index (*HumanDev*) developed by the Social Science Research Council as a measure of social wellbeing.[1] ~~This index measures levels of health, income, and education per CD, for the year 2015.~~ This index measures levels of health, income, and educational achievement per CD, for the year 2015. The index is based on a 10-point scale, with a score of 10 describing the highest possible state of human development. When ~~they are available~~possible, it is advisable to use already-published indices or measures in order to operationalize variables, as these indices have already been subject to the peer-review process, and can offer assurances that the operationalizations are acceptable translations of the constructs being studied. *HumanDev is* one such index, and I use it as the dependent variable in tests of H3.

Independent Variables:

Built Environment Investment (BEI): To measure investment in the built environment in H1a and H1b, I used a range of individual variables, summarized below :

total_permits: Total number of NYC building permits issued, ~~, over the period2013-2018,~~ per CD,

[1] https://measureofamerica.org/human-development/, accessed 30 May 2019.

Comentado [A6]: Which year(s) for this dataset?

Comentado [A7R6]: Period added.

Comentado [A8]: *Welfare?*

Comentado [A9R8]: I've changed "welfare" to "wellbeing" throughout

Comentado [A10]: Introduce this variable, year(s), type, per CD?

Comentado [A11R10]: Added.

Con formato: Fuente: Cursiva

Comentado [A12]: For all these variables, including depedent, independent, and control variables, I would suggest the author(s) to recheck these variables and clearly define them. Several issues:
1) Please add the time period for all the dependent variables/proxy variables.
2) when mentioning 2013-2018, does that mean the average annual, or the accumulated value? I guess the "total" means the sum of these six years, but what does "percentage of building permits" mean? Does that mean the average or sum from 2013-2018?
3) following the above comment, I think mixing accumulated values for a period and the average for a period might be acceptable to understand the correlation but it will be problematic when we calculate the correlation factors because of the different scales.
4) the variables across different time periods. Note that inconsistent time-period datasets may cause issues for internal vadility. For instance, if only #municipal wifi hotspots in 2018 is used, how can we involve this for a period between 2013-2018. In this case, the average wifi hotspots # between 2013-2018 would be used.

Comentado [A13R12]: 1. All updated.
2. For 2013-2018: I've indicated where it's the total over that period, or whether it's a percentage of that total.
3. Very good point. I have added a disclaimer.
4. I agree totally. This is clearly an issue in the study, but as this is a demonstration I felt it would be useful to illustrate a variety of ways to operationalize. I've made a note explaining this in the text.

Con formato: Fuente: Sin Negrita, Cursiva

Con formato: Fuente: Cursiva

Con formato: Fuente: Sin Negrita, Cursiva

Con formato: Sangría: Sangría francesa: 3,75 cm, Interlineado: sencillo

2013-2018

pct_permits_pub: Percentage of total NYC building ~~Percentage of building~~ permits issued to public agencies, ~~2013-2018,~~ per CD, 2013-2018

wifi_per10k: Total ~~n~~Number of m~~M~~unicipal wifi hotspots, per CD, 2018~~, per CD~~

sidewalkcaf_per10k: Total number of sidewalk café permits issued per 10,000 inhabitants, per CD, 2013-2018~~, per CD~~

pct_parks-rec: Percentage of total land dedicated to Parks and Recreation, per CD, 2018

pct_vacantland: Percentage of total land vacant, per CD, 2018

pct_cleanstreets: Percentage of city streets of "acceptable" cleanliness, NYC Dept. of Sanitation, per CD, 2018

Limitations in the available data have led me to include variables that do not all coincide in time. The variable measuring percentage of vacant land per CD, for example, was recorded for the year 2018, while several other variables reflect the period 2013-2018. In a scientific study, one would generally wish to avoid this inconsistency within the data set as it can threaten the internal validity of the study. That is, it can introduce confounding effects in the data that may make it difficult to know if observed cause-effect relationship is actually attributable to the variables being used. For the present demonstration, I have elected to include these variables because they allowed me to illustrate a variety of ways to operationalize variables and because the periods, while not identical, are nevertheless overlapping and therefore do describe at least in part a similar moment in time. ~~Readers should, however, be wary of doing so for purposes other than demonstration.~~

The above variables also present some challenges for interpreting the results, because~~as~~ they operate along different scales. The variable *pct_ vacantland*, for example, will only vary from 0 to 1, because it describes a percentage. The variable *total_ permits*, however,

Con formato: Sangría: Primera línea: 0 cm, Interlineado: sencillo

Con formato: Sangría: Sangría francesa: 3,75 cm, Interlineado: sencillo

Con formato: Interlineado: sencillo

Con formato: Sangría: Primera línea: 1,27 cm

Con formato: Fuente: Cursiva

might conceivably vary from 0 to infinity. When interpreting results involving these variables, researchers should therefore focus primarily on the sign and statistical significance of the coefficient, and not on the coefficient itself.

In tests of H2, I use the independent variable *ethnic_diversity* to measure the relative diversity of ethnic background within each CD~~,~~ in ~~2012-2016~~2010, per the six ethnicity categories as listed by the US Census Bureau.[2] I transformed ~~this~~ these raw census data with the Shannon index of population diversity (H), a statistical measure frequently used by biologists and others in the natural sciences to measure species diversity within a given population (see, for example. Worm, Lotze, Hillebrand, & Sommer, 2002). As I have used it in this study, this measure can be expressed with the formula:

$$H' = -\sum_{i=1}^{s} p_i \ln p_i$$

Where ***H'***=the value of Shannon index for the population being studied; ***s***=number of ethnic groups in the population, p_i=percentage of the total population made up by ethnic group **i**, and **ln**, the natural logarithm.

In tests of H3, I used change in environmental health (ΔEH) from 1995 to 2015, per CD, as an independent variable. Previously, in~~n~~ H2 and H3, I have used the proxy variables *child_asthma* and life_expect to stand in for EH. These proxies, however, do not make it possible to measure a change in environmental quality over time, as they are based on a single year of observation, and data on these variables were not available for other years. ~~no further data was available.~~ For that reason, I use a different operationalization for~~for~~ ΔEH, as this variable requires at least two obsevations in order to describe a change over time. I therefore

Comentado [A14]: I am a bit confused why to use the tree census to measure the environmental health. This variable was the dependent variable in the other two hypotheses and measured by the two proxy variables. Is that clearer to change the variable name?

Comentado [A15R14]: I've included an explanation and changed the variable name.

Con formato: Fuente: Cursiva

[2] Per the United States Census information used in this study, racial and ethnic identity was classified according to the following categories: "Asian/Pacific Islander", "Black","Latino","Native American", "White", and "Other".

~~In order to operationalize this variable, I~~ used data from the 1995 and 2015 NYC Tree Census. I used the Tree Census data to calculate the percentage of healthy trees in each CD in 1995 (*pct_healthytrees95*) and in 2015 (*pct_healthytrees15*). The independent variable ΔEH~~*Δhealthytrees_1595*~~, which measures the change in the percentage of healthy trees in a given CD, can thus be expressed as *pct_healthytrees15- pct_healthytrees95.*

Control Variables~~:~~

Con formato: Fuente: Sin Negrita, Cursiva

New York is a city with great economic and social inequality among Community Districts, and intuitively these differences among CDs are likely to play a role in the study. I take this into account by including several control variables in the regression models, while also controlling for a range of other factors that may account for some variation across CDs, but which are not the object of the present study.

income:	Median household income, per CD, 2018~~, per CD~~
degree:	Percentage of adults with at least an undergraduate degree, ~~2018,~~ per CD, 2018
airpollution:	Fine particulate matter~~,~~: annual ~~2017~~ mean fpm in mcg, per m^3, per CD, 2017
derelict_towed:	Total number of derelict vehicles towed, per CD, 2013-2018~~, per CD~~

Methods

Con formato: Fuente: Sin Cursiva

As described in the previous section, this~~This~~ demonstration involves many variables, and it is generally advisable to reduce the number of variables involved in a regression model, especially if these a~~variables a~~re related in some way. An Exploratory Factor Analysis (EFA) is used to simplify and consolidate variables, and can help make regression models more robust (Lattin, Carroll, & Green, 2003). If several variables in the study are, in reality, measuring related phenomena, a factor analysis can help determine which variables are related, and to

Con formato: Color de fuente: Automático

what degree this is so. This would allow the researcher to reduce the number of individual variables in the model by grouping several variables into a single factor variable. In this demonstration, I detect a relatively high correlation among several variables related to investment in the built environment. This correlation suggests that it may be useful to perform an exploratory factor analysis (EFA).

As factor analysis requires both statistical testing and the individual interpretation of the researcher, it involves a degree of subjectivity. The researcher must decide how many, if any, of the new, composite variables make sense and are of use. The researcher also must determine what these new groupings of variables are actually measuring. If, for example, a factor analysis suggested that we might group the variables *airpollution*, *pct_healthytrees*, and *child_athsma*, then we might logically retain that factor and describe it as a compound measure of environmental quality. A factor variable grouping together *wifi_per10k*, *life_expect*, *derelict_towed*, variables that do not share an immediately discernible link among them, might be somewhat more difficult to justify as a logical grouping of variables, and could therefore be discarded.

After performing a factor analysis and varimax rotation—a transformation of the data used to strengthen and clarify the relationship among the variables—one can study the specific variables being grouped into composite factor variables, and assess which, if any, might be retained. Results are as follows:

<<insert Table 1 about here>>

As a rule of thumb, factors with Eigenvalues above 1.0 should be maintained. As we see above, only factors 1 and 2 reach that level, and together they account for 62.01% of the variance in the data, that is, one could say they provide about two thirds of the explanation for why the

Con formato: Color de fuente: Automático

Con formato: Color de fuente: Automático

data are as they are[3]. The factor loading table helps us understand which variables should be grouped together to form factor variables. In general, variables with factor loadings above 0.5 should be maintained. The rest should not be included as part of the factor. By that logic, we would have two groupings: the first, Factor 1, above, would group the variables *total_permits*, *wifi_per10k, and sidewalkcafes_per10k*. One could certainly argue that the total number of building permits issued, the total wifi hotspots, and sidewalk café applications might all speak of investment in the built environment. I thus accept this grouping as a factor variable and combine the three individual variables into a new, composite factor variable, which I label *Built Environment Investment (BEI).*

The second grouping, per the EFA, would include the variables *pct_cleanstreets* and *pct_vacantland*. The variable *pct_parks-rec*, however, does not reach the .50 threshold and is not assigned to either factor. This second factor is less clear than the first. While there is a statistical linkage between these variables, its meaning is not immediately clear, and is not necessarily consonant with the theoretical framework used in this demonstration. I can therefore discard the factor. I consider it more useful to maintain the individual variables as control variables than to synthesize them into a factor variable with unclear meaning.

Regression and Logistic Regression

For tests of H1 and H2, I performed ~~a~~ multiple linear regressions to explain the effect of Built Environmental Investment~~El~~ (BEI) on Environmental Health (EH) as well as the effect of Ethnic Diversity on Life Expectancy (Kellermanns & Eddleston, 2006). In general, a linear regression equation estimates a complex linear model to describe the relationship between the dependent and independent variables, controlling for the other parameters included in the

[3] A ~~S~~scree plot~~s~~ of Eigenvalues can offer a useful visual tool for deciding which factors to be retained.~~included~~.

Con formato: Color de fuente: Automático

Con formato: Fuente: Sin Negrita

Con formato: Fuente: Sin Negrita

Con formato: Color de fuente: Automático

model. The basic equation for a line, *y=mx+b*, is, in fact, a very simple linea regression model, providing an estimation of how a change in x will affect the value of y. If, for example the slope of this line, *m*, is positive, then increased values of x will result in increased values of y. The regression models in this ~~e case study~~ example are more complex than the simple equation for a line, but the principle is the same. I used Logistic Regression to test H3. Logistic regression describes the relationship between a binary dependent variable (in this case, Community Districts that experienced a decline in tree health from 1995 to 2015 versus those that did not) and one or more independent variables (Palmer, Jennings, & Zhou, 1993; Sherer & Lee, 2002).

Con formato: Fuente: Sin Negrita, Sin Cursiva

Descriptive Statistics ~~, FACTOR ANALYSIS,~~ And Results

Con formato: Fuente: Sin Cursiva

Before presenting the results of the hypothesis tests, researchers should provide descriptive statistics regarding the data being used. These ~~tests~~ help the reader understand the general trends in the data (means, standard deviation, maximum and minimum values of each variable, for example), and can also help verify that the variables being used are, in fact, measuring different phenomena and are not excessively correlated among themselves. This information can be communicated through a correlation table (Table ~~1~~2), which in this case has been produced with the STATA 14 application.

~~application.~~

Although the correlation table does not constitute a test of any of the hypotheses used in the study, it can alert us to potential instances of collinearity among our variables; if two variables that we assume to be independent are, in fact, very highly correlated, then they may distort our regression models and compromise our results. As a rule of thumb, very high

correlations (over .80 or .90) are considered to present a risk of collinearity of variables[4]. ~~5~~[5]~~5.~~

Table ~~1~~ 2 displays the correlations and descriptive statistics.

<<insert Table ~~1~~ 2 about here>>

We observe a relatively high correlation between the variables *wifi_per10k* and *airpollution* (0.8042), which could likely be explained by the general tendency to place public wifi hotspots in areas with high levels of automobile traffic, particularly in Manhattan. As this value is quite close to the 0.80 threshold, I include it in the model. Higher levels of correlation can be observed between the variables *income* and *human_dev* (0.8714) as well as between the variables *degree* and *human_dev* (0.8935) and *income* and *degree* (0.8961). Although I have elected to include these variables in the regression models, more study would be needed to verify that the correlation is not weakening the regression models I use below.

~~**Factor Analysis**~~

~~An Exploratory Factor Analysis (EFA) is used to simplify and consolidate variables, and can help make regression models more robust (Lattin, Carroll, & Green, 2003). If several variables in the study are, in reality, measuring related phenomena, a factor analysis can help determine which variables are related, and to what degree this is so. This would allow the researcher to reduce the number of individual variables in the model by grouping several variables into a single factor variable. In this demonstration, I detect a relatively high correlation among several variables related to investment in the built environment. This correlation suggests that it may be useful to perform an exploratory factor analysis (EFA).~~

Con formato: Color de fuente: Fondo 1

[4] A useful tool for studying the effects of collinearity is the variance inflation factor (VIF), which estimates the degree to which the observed collinearity affects the variance of a given variable in the model being tested.

[5]

As factor analysis requires both statistical testing and the individual interpretation of the researcher, it involves a degree of subjectivity. The researcher must decide how many, if any, of the new, composite variables make sense and are of use. The researcher also must determine what these new groupings of variables are actually measuring. If, for example, a factor analysis suggested that we might group the variables *airpollution*, *pct_healthytrees*, and *child_athsma*, then we might logically retain that factor and describe it as a compound measure of environmental quality. A factor variable grouping together *wifi_per10k*, *life_expect*, *derelict_towed*, variables that do not share an immediately discernible link among them, might be somewhat more difficult to justify as a logical grouping of variables, and could therefore be discarded.

After performing a factor analysis and varimax rotation—a transformation of the data used to strengthen and clarify the relationship among the variables—one can to study the specific variables being grouped into composite factor variables, and assess which, if any, might be retained. Results are as follows:

<<insert Table 2 about here>>

As a rule of thumb, factors with Eigenvalues above 1.0 should be maintained. As we see above, only factors 1 and 2 reach that level, and together they account for 62.01% of the variance in the data, that is, one could say, they provide about two thirds of the explanation for why the data is are as it isthey are. The factor loading table helps us understand which variables should be grouped together to form factor variables. In general, variables with factor loadings above 0.5 should be maintained. The rest should not be included as part of the factor. By that logic, we would have two groupings: the first, Factor 1, above, would group the variables *total_permits*, *wifi_per10k*, *and sidewalkcafes_per10k*. One could certainly argue that the total number of building permits issued, the total wifi hotspots, and sidewalk café applications might all speak of investment in the built environment. I thus accept this grouping as a factor

~~variable and combine the three individual variables into a new, composite factor variable, which I label *Built Environment Investment (BEI).*~~

~~The second grouping, per the EFA, would include the variables *pct_cleanstreets* and *pct_vacantland*. The variable *pct_parks-rec*, however, does not reach the .50 threshold and is not assigned to either factor. This second factor is less clear than the first. While there is a statistical linkage between these variables, its meaning is not immediately clear, and is not necessarily consonant with the theoretical framework used in this demonstration. I can therefore discard the factor. I consider it more useful to maintain the individual variables as control variables than to synthesize them into a factor variable with unclear meaning.~~

Results

Table 3 summarizes the results of the tests of H1a and H1b. For the sake of clarity, I have substituted the variable names for simple variable descriptions in the tables.

~~I performed a multiple linear regression to explain the effect of BEI on EH (Kellermanns & Eddleston, 2006). In general, a linear regression equation estimates a complex linear model to describe the relationship between the dependent and independent variables, controlling for the other parameters included in the model. The basic equation for a line, *y=mx+b*, is, in fact, a very simple linea regression model, providing an estimation of how a change in x will affect the value of y. If, for example the slope of this line, *m*, is positive, then increased values of x will result in increased values of y. The regression models in the case study example are more complex than the simple equation for a line, but the principle is the same.~~ The results are as follows:

<<insert Table 3 about here>>

~~For the sake of clarity, I have substituted the variable names for simple variable descriptions in the tables.~~

Con formato: Color de fuente: Fondo 1

Comentado [A16]: I didn't see the results in the table, but for linear regression model, in addition to P value, the (adjusted) R square should be reported as well.

Comentado [A17R16]: I have now included the R squared in the tables, for all but the logistic regression, for which R-squared cannot be calculated. I would have to use a pseudo R-squared. As this isn't really comparable to R-squared as used in the other models, I would prefer to omit for models 7 and 8. Given space constraints I have only included the statistic in the table and have not explained it. I can add this without problem, but it may be too much information for this rather introductory demo.

In tests of H1a and H1b, I used childhood asthma rates as a proxy for EH. Model 1 includes only the control variables and shows a statistically significant and negative relationship between childhood asthma and derelict vehicles towed, vacant land, street cleanliness, life expectancy, and percentage of adults with at least bachelor's degree. This is to say, for example, the study suggests that for every additional unit of acceptable street cleanliness in a community, we would expect childhood asthma rates to decrease. The clearest and most robust relationship here is the link between life expectancy and childhood asthma rates. This strong correlation appears in all models in Table 3, even when the independent and interaction variables are added in Models 2 and 3. Perhaps unsurprisingly, the results also show a significant and positive relationship between air pollution and childhood asthma rates.

Model 2 introduces the independent variable, a compound index measuring BEI, as developed through the exploratory~~previous~~ factor analysis. The results describe a modest, but statistically significant negative relationship between BEI and childhood asthma. That is, as investment in the built environment increases, childhood asthma rates decline modestly. It should be noted that these results have a p-value between 0.05 and 0.10, at the upper limit of what social scientists would consider statistically significant. A p-value, which ranges from 0 to 1, measures the strength of the evidence to reject what is called the "null hypothesis" or H_0. That is, if the hypothesis predicts that BEI and childhood asthma rates *are* intercorrelated, the "null hypothesis" would be that these variables are *not*, in fact intercorrelated. A p-value approaching 0.10 suggests that the probability that we would observe data like ours if the null hypothesis were true is nearly 10%. This means, colloquially speaking, that when one observes results such as these, it might lead us to make a "false positive" judgement between 5 to 10% of the time. In many cases, social science researchers set the threshold at 5%. ~~I~~~~We~~ ~~nevertheless~~One might therefore be inclined to ~~can~~ report finding support for H1a, with the caveat of a relatively high, but nevertheless marginally acceptable, p-value. Once further

Comentado [A18]: Again, I didn't see the generated model, but P value should be for the coefficient. If it is for y=ax+b, we generally ignore the p-value for the intercept. Meanwhile, to evaluate the overall fit of a linear model, the R-squared value should be reported.

Comentado [A19R18]: Please see above: the coefficient p-values are indicated at levels of significance. I would be happy to include them, but this is rarely, if ever, done in social science tables.

Comentado [A20]: The ET finds this very helpful in reading the tables. Could such a thing be written for the numbers under each model?

Comentado [A21R20]: This is indicated in the tables --I have followed the standard format for journals in the social sciences. Exact P-values for the coefficients aren't generally reported, but instead are marked with +, *,**, or *** to indicate the different p-values for those results that are statistically significant at the p<0.10, p<0.05, p<0.01, an p<0.001 levels. I have calculated all the p-values, but since this is meant to be a demonstration of how some Social Science methods might be used in architecture, it seems like it would be best to follow the standard table format.

variables are added in model 3, however, even this very modest correlation is no longer evident, and we would as a result report no support for H1a.

Model 3 adds as an interaction term the percentage of building permits issued to clients in the public sector, to test whether greater proportion of public investment would influence the degree to which BEI affected the childhood asthma rate. Adding an interaction term allows us to estimate the effect of an individual variable as well as the effect of that variable in combination with another. We find no statistically significant relationship, neither for the variable on its own, nor as an interaction term, and therefore do not find support for H1b.

Table 4 contains the results of tests of H2, which predicted a positive relationship between Environmental Health and Ethnic Diversity. In these models I have used life expectancy as a proxy for EH, assuming—perhaps debatably—that communities with the healthiest environments would also have the longest life expectancy. Results are as follows:

<<insert Table 4 about here>>

Model 4 contains the control variables only, and it shows that few of the control variables are significant moderators of life expectancy, though we do see a positive and significant relationship between the Human Development Index score and Life Expectancy, as one might expect, and a less robust, but nevertheless significant and negative relationship between vacant land and life expectancy. That is, as the percentage of vacant land in a community increases, life expectancy decreases.

Model 5 introduces a per capita measurement of building permit applications to the control variables, as a check of the robustness of the H1a tests. Here too, I find a positive and statistically significant result (at the 0.10 level), providing further support for H1a.

Comentado [A22]: This can be demonstrated by the P-value for the coefficient. Please add that.

Comentado [A23R22]: Please see above: p-values have been reported at significance levels, per standard social science practice. The positive relationship, however, would not be described by the p-value, but rather by the sign of the coefficient. As the p-value I report in model 6 is < 0.05 (here, denoted with a *), and the sign of the coefficient is positive, I observe a positive relationship, as predicted.

In model 6, I test the relationship between ethnic diversity, as operationalized with the Shannon measure of population diversity, and life expectancy. The model shows a positive and statistically significant result, providing support for H2.

Hypothesis 3 predicted lower levels of social wellbeing for those CDs that had experienced a decline in EH over the period 1995-2015. I tested this hypothesis through a Logistic Regression. ~~Logistic regression describes the relationship between a binary dependent variable (in this case, CDs that experienced a decline in tree health from 1995 to 2015 versus those that did not) and one or more independent variables (Palmer, Jennings, & Zhou, 1993; Sherer & Lee, 2002).~~ In this case, the ~~in~~dependent variable was *human_dev*, the index value for human development. The results are reported in Table 5:

Comentado [A24]: This is depedent variable in your design, right?

Comentado [A25R24]: That's right. I have corrected.

<<insert Table 5 about here>>

~~Model 7 includes the control variables only, and Model 8 introduces the independent variable *human_dev*.~~ T~~Before proceeding with the description of these results, it must first be noted that t~~he p-value for the ~~Model 5 test is at the upper limit end of acceptable statistical significance (p=0.0689) and the~~ Model ~~6~~ 7 test is ~~not statistically~~beyond the upper limit of acceptable statistical significance ~~significant at all~~ (p=0.1474; p-values less than 0.05 or 0.10 are generally understood to be statisitically significant). It would be inadvisable to consider as significant any results from ~~Model 7~~this model, and we therefore~~, and those in Model 8 should be reported with some skepticism. Model 7 shows a modest negative relationship between the control variable BEI and decline in tree health, and an extremely modest, positive and statistically significant relationship between derelict vehicles towed and the likelihood that a CD had experienced a decline in tree health. As model 8 is not statistically significant, however, we~~ report no support for H3. I summarize the hypothesis tests in Table 6, below:

<<insert Table 6 about here>>

~~5:~~ DISCUSSION

As I have illustrated above, the results provide relatively modest support for ~~hypotheses 1a~~ hypothesis~~and~~ 2. As this demonstration study did not include a theoretical discussion of *why* I had predicted this relationship of variables, one can, for the moment, only speak of correlation; I have not provided sufficient theoretical backing to argue for any potential causal relationship among variables. Indeed, there are many other possible alternative explanations for the findings here. For example, one could argue that ~~for~~ even if we *had* found support for H1a, the potential linkage of BEI and childhood asthma rates might, in fact, be relatively unsurprising: areas with greater building activity might, in general, be those with higher median incomes, and those communities with higher median incomes, one might expect, could in general benefit from better environmental health. Seen in this way, the results might instead be interpreted as support for the proposition that median household income, more than investment in the built environment, actually determines the health of a given community environment. The lack of support for H1b lends further credence to this idea.

Con formato: Fuente: Cursiva

One could also propose alternative explanations for the positive correlation of ethnic diversity in a community and life expectancy, arguing once again that income is a more relevant determinant of life expectancy, and that ethnic diversity is, in fact, indirectly linked to the median income of a given CD. The correlation table, for example, shows a much more robust correlation between income and life expectancy (0.5939) than between ethnic diversity and life expectancy (0.229). This may indeed be the case, but it is nevertheless worth noting that the three most diverse Community Districts in New York, the 8th, 9th, and 10th districts of Queens, with median annual household incomes around $70,000, have greater average life expectancies (between 84 and 85 years) than do the 5th and 6th districts of Manhattan, with median annual incomes of over $140,000 and diversity rankings of 43rd and 18th, respectively.

As the above may suggest, the data themselves are unlikely to present a coherent or decisive case for causality or against alternative explanations. It is the researcher's obligation

to make a convincing theoretical argument for a given causal relationship, and, either explicitly or implicitly, against the alternative explanations. The key is to situate both the hypotheses and the findings within the existing literature and to develop a coherent theoretical case for a generalizable, causal relationship among the variables. This, of course, requires an extensive body of previous empirical and theoretical work, one which may not always be available in our field. We may at times need to venture into the literatures of other disciplines, and borrow extensively but responsibly. Once we navigate the occasionally intimidating jargon of other fields, there is little to fear in these waters, and much to gain. Also, as the culture of citation in architecture is still relatively young and less robust than in the social sciences[6], there remains much work to be done to strengthen and enrich the conversation we researchers are having with one another in the pages of journals such as this one.

There are, of course, limitations to the approach outlined here, especially if one considers the broader discussion about the nature and role of knowledge, and of research itself. A deductive, quantitative approach, such as the one used here, tends to reinforce a positivist worldview, in which only the measurable is knowable, and in which a single, objective truth is presumed to exist. In this understanding of science, the researcher him or herself would act as a neutral filter, allowing the natural truths of the world to emerge through empirical tests. A post-positivist or interpretive approach, meanwhile, assumes that knowledge is inevitably partial or imperfect, and that the researcher is constrained and conditioned by her or his cognitive biases and interpretations of the world. Research

[6] The relative weakness of the culture of citation in architecture is evident in a comparison of the SCIMAGO impact factor rankings for journals in management and those in architecture. Impact factor measures the average number of citations per article published in a given journal. The top three management journals, per the SCIMAGO rankings, are the *Academy of Management Journal*, the *Academy of Management Review*, and the *Brookings Papers on Economic Activity*, with impact factors of 8.548, 7.880, and 6.851, respectively. The top three architecture journals, per these rankings, are *Research in Engineering Design, Journal of Building Performance Simulation*, and *Design Studies*, with impact factors of 1.024, 0.957, and 0.941, respectively.

conducted from this standpoint would generally employ an inductive and qualitative approach and would acknowledge or embrace the subjective role of the researcher in the research process.

In architecture research, I believe that quantitative measures and methods like those demonstrated here might be useful, no matter one's position in the epistemological debate outlined above, because these measures may be considered merely one aspect of our understanding of a phenomenon, to be complemented with other qualitative or inductive approaches. Broadening in this way the range of tools available for the architecture researcher may, I hope, enrich the scholarly conversation and debate. I believe that quantitative methods used in the social sciences may help fuel that conversation, and hope that this article may serve as a useful tool for those interested in taking a step in that direction.

REFERENCES:

Agresti, A. (2018). *Statistical Methods for the Social Sciences*. Harlow, UK: Pearson.

Amburgey, T. L., & Dacin, M. T. (1994). As the left foot follows the right? The dynamics of strategic and structural change. *Academy of Management Journal*, *37*(6), 1427–1452.

Bono, J. E., & McNamara, G. (2011). From the editors. Publishing in AMJ - Part 2: Research design. *Academy of Management Journal*, *54*(4), 657–660.

Carroll, G. R., & Delacroix, J. (1982). Organizational mortality in the newspaper industries of Argentina and Ireland: An ecological approach. *Administrative Science Quarterly*, *27*(2), 169–198.

Cohen, M. D., March, J. G., & Olsen, J. P. (1972). A garbage can model of organizational choice. *Administrative Science Quarterly*, *17*(1)

Davis, M. S. (1971). That's interesting! Towards a phenomenology of sociology and a sociology of phenomenology. *Philosophy of the Social Sciences*, 1(4), 309–344.

Dunn, M. B., & Jones, C. (2010). Institutional logics and institutional pluralism: The contestation of care and science logics in medical education, 1967-2005. *Administrative Science Quarterly*, (55), 114–149.

Glynn, M. A., & Abzug, R. (2002). Institutionalizing identity: Symbolic isomorphism and organizational names. *Academy of Management Journal*, *45*(1), 267–280.

Hannan, M. T., & Freeman, J. (1977). The population ecology of organizations. *American Journal of Sociology*, *82*(5), 929.

Kellermanns, F. W., & Eddleston, K. A. (2006). Corporate entrepreneurship in family firms : A family perspective. *Entrepreneurship Theory and Practice*, (662), 809–831.

Lattin, J., Carroll, J. D., & Green, P. E. (2003). *Analyzing Multivariate Data*. Belmont, CA: Brooks/Cole.

Lawrence, P. R., & Lorsch, J. W. (1967). Differentiation and integration in complex organizations. *Administrative Science Quarterly*, *12*(1), 1–47.

Miller, D. C. & Salkind, N. J. (2002). *Handbook of Research Design & Social Measurement-6th Edition*. Thousand Oaks, CA: Sage Publications.

Palmer, D. A., Jennings, P. D., & Zhou, X. (1993). Late adoption of the form by multidivisional large U. S. corporations: Political , institutional , and economic sccounts. *Administrative Science Quarterly*, *38*, 100–131.

Romanelli, E. (1989). Environments and strategies of organization start-up: Effects on early survival. *Administrative Science Quarterly*, *34*(3), 369–387.

Sherer, P. D., & Lee, K. (2002). Institutional change in large law firms : A resource dependency

Con formato: Superíndice

and institutional perspective. *Academy of Managemant Journal*, *45*(1), 102–119.

Simon, H. (1991). Bounded rationality and organizational learning. *Organization Science*, *2*(1), 125–135.

Sparrowe, R., & Mayer, K. (2011). Publishing in AMJ — Part 4: Grounding Hypotheses. *Academy of Management Journal*, *54*(6), 1098–1102.

Stinchcombe, A. (1959). Bureaucratic and craft administration of production: A comparative study. *Administrative Science Quarterly*.

Thornton, P., & Ocasio, W. (1999). Institutional logics and the historical contingency of power in organizations: Executive succession in the higher education publishing industry, 1958-1999. *The American Journal of Sociology*, Vol. 105, pp. 801–843.

Trochim, W. M. K., & Donnelly, J. P. (2008). *The Research Methods Knowledge Base* (3rd ed.). Mason, OH: Cenage.

Worm, B., Lotze, H. K., Hillebrand, H., & Sommer, U. (2002). Consumer versus resource control of species diversity and ecosystem functioning. *Nature*, *417*(June), 848–851.