Regression Analysis

Quantitative exploration of interactions between the built environment and spatial behaviour

Romein, A.; van Rijn, Susanne

Publication date
2022

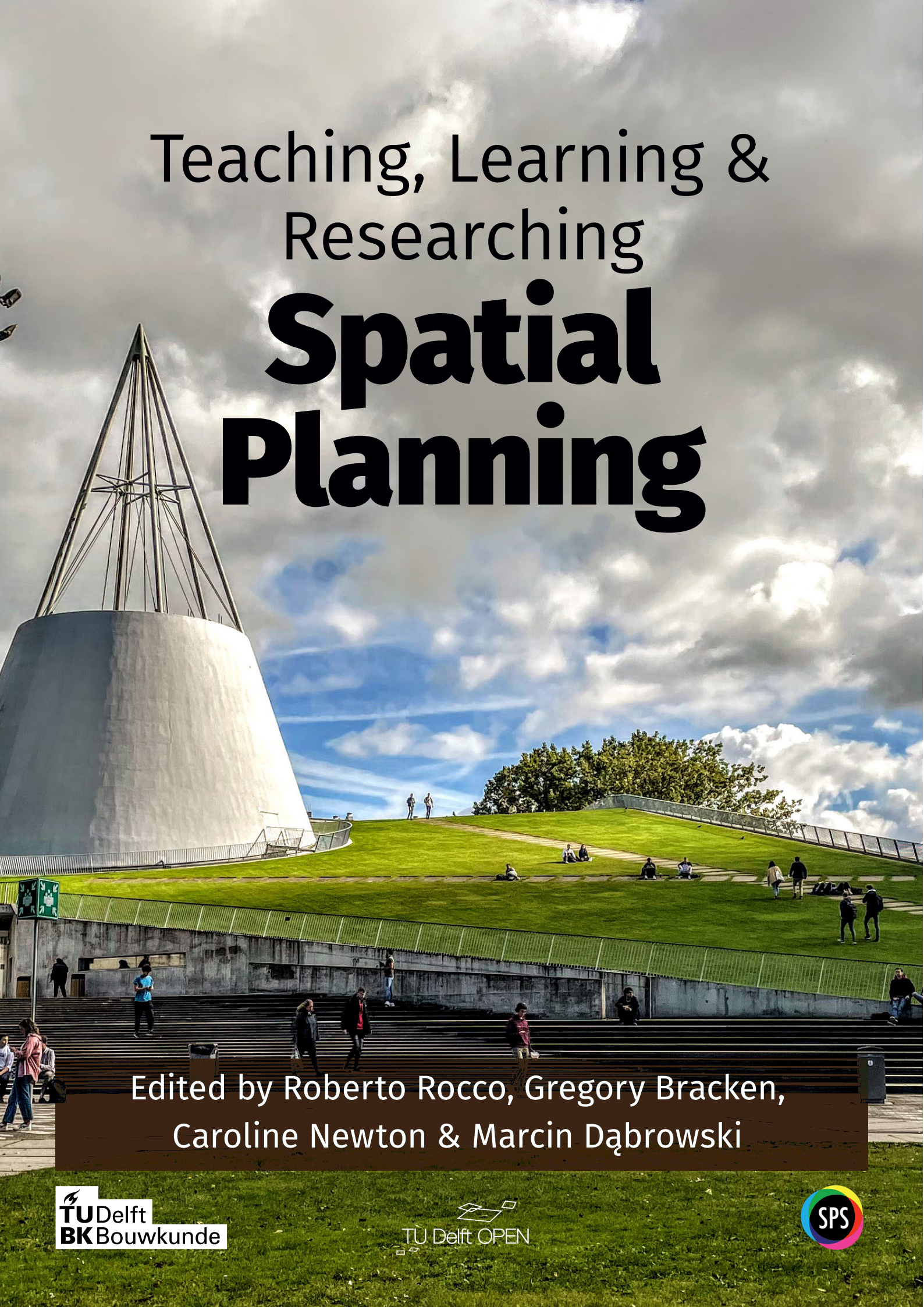
Document Version
Final published version

Published in
Teaching, Learning & Researching Spatial Planning

Citation (APA)
Romein, A., & van Rijn, S. (2022). Regression Analysis: Quantitative exploration of interactions between the built environment and spatial behaviour. In R. Rocco, G. Bracken, C. Newton, & M. Dabrowski (Eds.), *Teaching, Learning & Researching Spatial Planning* (pp. 246-262). TU Delft OPEN Publishing.

Teaching, Learning & Researching

Spatial Planning



Edited by Roberto Rocco, Gregory Bracken, Caroline Newton & Marcin Dąbrowski

TU Delft BK Bouwkunde

TU Delft OPEN

SPS

Teaching, Learning & Researching
Spatial Planning

TOOLS, CONCEPTS AND IDEAS TAUGHT AT THE SECTION OF SPATIAL PLANNING AND STRATEGY OF THE
OF URBANISM, FACULTY OF ARCHITECTURE AND THE BUILT ENVIRONMENT
DELFT UNIVERSITY OF TECHNOLOGY, THE NETHERLANDS.

Published by
TU DELFT OPEN

Edited by
ROBERTO ROCCO, GREGORY BRACKEN, CAROLINE NEWTON & MARCIN DĄBROWSKI

Design and layout
ROBERTO ROCCO

Language review & copy editing
GREGORY BRACKEN

Contact
SECTION SPATIAL PLANNING & STRATEGY, DEPARTMENT OF URBANISM
FACULTY OF ARCHITECTURE AND THE BUILT ENVIRONMENT, DELFT UNIVERSITY OF TECHNOLOGY
JULIANALAAN 134, 2628 BL, DELFT, THE NETHERLANDS
ENQUIRIES: KARIN VISSER , E-MAIL: SPATIALPLANNING-BK@TUDELFT.NL
ISBN/EAN: 978-94-6366-604-6
https://doi.org/10.34641/mg.50

COVER: TU DELFT CENTRAL LIBRARY BY MECANOO ARCHITECTS, TU DELFT. PHOTO BY R. ROCCO (2019).

TU Delft BK Bouwkunde

TU Delft OPEN

URBANISM

Regression Analysis

Quantitative exploration of interactions between the built environment and spatial behaviour

ARIE ROMEIN

RESEARCHER URBAN AND REGIONAL DEVELOPMENT, TU DELFT, AROMEIN109@HETNET.NL

SUSANNE VAN RIJN

CONSULTANT & RESEARCHER AT SPEELPLAN, ALMERE, SEMVANRIJN@GMAIL.COM

In many graduation projects in Spatial Planning and Strategy (SP&S), empirical research and spatial design are intertwined. This chapter introduces regression analysis; a 'family' of related models of quantitative statistical analysis in empirical research. It is very appropriate to study interactions between the built urban environment and people's spatial behaviour in a project location at a high level of quantitative precision. The outcomes of such quantitative studies can be very useful in urban design or planning, either in preliminary empirical research, i.e. prior to design, or in the iterative cycles of Research by Design. There are several types of regression models; which one is most appropriate for your project (if any) depends on the specific questions about that interaction you want to answer, and on the empirical data that is available for that answer.

REGRESSION ANALYSIS, QUANTITATIVE METHODS, SPATIAL BEHAVIOUR, SPATIAL QUALITIES

1. Introduction

As a student (or practitioner) in Spatial Planning & Strategy (SP&S), you might plan to study the interactions of the built environment and people's spatial behaviour in an urban area. Most students use a broad kaleidoscope of methods to collect and analyse data on these interactions, including literature review, interviews, on-site observations, policy analysis, case studies, and mapping. With the aim of 'deep understanding' of processes that influence these interactions in their project areas, they usually assess obtained data at face value in a qualitative manner.

For sure, these analyses yield insights for interventions in built environments to better adjust their qualities to people's behaviour. Great! that is what spatial planning or urban design graduation projects are about! But these analyses rarely lead to detailed insights on the variety of spatial behaviour by different groups of (potential) users in that built environment. Not uncommonly, there is only some implicit notion of an undefined 'average user'. Hence, spatial interventions based on such implicit notions are not optimal from the perspectives of most users.

This chapter presents various models of a specific type of statistical technique, regression analysis, that you can use to study spatial behaviour by users of the built environment, at a high level of quantitative precision. This behaviour can be either revealed or stated, i.e. being either actually displayed or planned intentional behaviour. To study these different types of behaviour, regression analysis can fit in different stages of your project; either in empirical research prior to design or in the iterative cycles of research by design. Regression analysis is being used, yet in many different academic fields it appears rather unknown among graduates (and even practitioners) of SP&S. That is a missed opportunity because it can yield key knowledge that you can use for appropriate spatial interventions.

This chapter starts with sections two and three that present elementary features of the data that is required for regression analysis. Next, sections four and five discuss two key principles of regression analysis: 1) it explores causal relations, such as qualities of the built environment as explanation of people's spatial behaviour, and 2) it reaches conclusions on these causal relations for an entire population based on analysis of only a subsection (sample) of cases from that population. If you are primarily interested in the types of research findings that regression analysis provides you can skip sections two to five. Section six presents the simplest model to explain the essence of regression analysis. Next, sections seven to nine present more advanced models by means of examples of their application taken from literature. If you are afraid of the word 'statistics', do not worry, there is no reason to fear the word. Any mathematical explanations as to the basis of these models is limited to a basic minimum. Finally, section ten summarises in a very general manner the practical usefulness of regression analysis for spatial interventions by urban design or planning. But in spite of its usefulness, it ends with the conclusion that regression analysis is just one among various methods, including qualitative ones, that are required in complex multidimensional graduation projects.

Case ID	Age ratio	Gender nominal	Education ordinal	Place of Residence nominal	Frequency of Visits ratio
1	12	1	1	2	4
2	23	2	2	1	2
3	47	2	2	1	5
4	21	2	2	1	1
5	34	1	3	2	0
6	25	1	1	3	4
7	60	2	3	1	6
8	18	1	2	1	2
9	57	1	2	1	5
10	25	1	3	1	3
11	35	2	1	2	1
12	42	2	2	3	0
13	15	1	1	2	0
14	52	1	3	2	4
15	29	2	3	1	2
...	..	.	.	.	.

Table 1: Structure of the SPSS dataset.

2. Database

Regression analysis requires an adequate quantitative database. This database has the form of a matrix: it consists of rows, columns, and cells (see **Table 1**). These three elements correspond to the fundamentals of a database: cases, variables, and values. *Cases* are the individual members of the population in your project that are in the sample. They represent the *unit of research* of the project. In SP&S graduation projects, the unit of research is mostly either built-up areas in the city – for instance postal code areas – where spatial behaviour takes place, or the individuals who practice spatial behaviour in these areas.

The *variables* in the database are the features of the unit of research that are relevant in the project. In the case of built-up areas, variables can be building density, building typologies, population size, available amenities, or level of liveability. In the case of individuals as units of research, variables can be socio-demographics like age, income, and educational level, use of amenities at specified hours, or frequency of visiting the postal code areas in the project location. Finally, a *cell* is the intersection of a specific column and row that contains the *value* of the variable in the column measured for the case in the row.

Statistical analyses like regression are mathematical operations that require numerical data. Therefore, all values in the database are encoded by a number even when the corresponding variable is not numerical in nature. For example, the codes of gender in **Table 1** (names of variables in this chapter are in capital letters) are 1 = male and 2 = female. The mathematical operations are carried out by statistical software. A rather user friendly and comprehensive package that is often used by researchers in social sciences, SPSS, is useful for SP&S graduation projects.

	Nominal categorical	**Ordinal** categorical	**Interval** continuous	**Ratio** continuous
Order of values	No	Yes	Yes	Yes
Distances between values	No	No	Yes	Yes
Zero point	No	Yes/No	No	Yes
Examples	Gender Province of residence	Type of education Spatial entity	Temperature	Population size Areal surface Annual Income

Table 2: Scale of measurement of variables.

3. Types of variables

One essential feature of the variables that are included in the analysis – their *scale of measurement* – is critical for the choice of the appropriate regression model. There are four different scales of measurement: nominal, ordinal, interval, and ratio. They differ in three aspects: 1) the indication of order (ranking) of values, 2) distances between points on the value scale, and 3) meaning of the zero point on that scale (see **Table 2**). The scale of measurement of each variable in **Table 1** is mentioned at the top of its column.

Nominal variables are *categorical*. The values of nominal variables take on only a few possible discrete categories in words or names. Examples in **Table 1** are gender and province of residence in the Netherlands, e.g. Groningen, Zeeland, or South Holland. There is no order of these values in terms of large versus small or more versus less. Further, nominal values have no unit of measurement, like € or $ of the variable annual income. Without a unit of measurement, calculations with values are not possible and, hence, quantitative distances between values cannot be computed. Finally, the value scale of nominal variables has no zero point. After all, that would mean that a case (person) has no gender or lives in no province.

Ordinal variables are also *categorical*. But unlike nominal variables, the values of ordinal variables have an indication of order: a secondary level of education is higher than a primary but lower than a tertiary level **(Table 1)**. But distances between these values cannot be computed in numerical terms. Some ordinal variables have a zero point (like Level of Education) but most have not. A common type of ordinal variable is measured on a Likert scale, for instance, a five-point scale on which people score their satisfaction with quality of public space in their residential neighbourhood: 1) very dissatisfied, 2) dissatisfied, 3) neutral, 4) satisfied, 5) very satisfied. Mind that neutral is not a zero point!

Whereas nominal and ordinal variables have to be encoded by numbers to be included in statistical analysis, *interval* and *ratio* variables are *numerical* by nature. There is a difference between these two types **(Table 2)** but that is irrelevant for the choice of the appropriate statistical technique. Therefore, SPSS does not distinguish between them and takes them together as *scale variables*. Scale variables are *continuous*: whereas nominal and ordinal variables take on only a few discrete values, interval and ratio

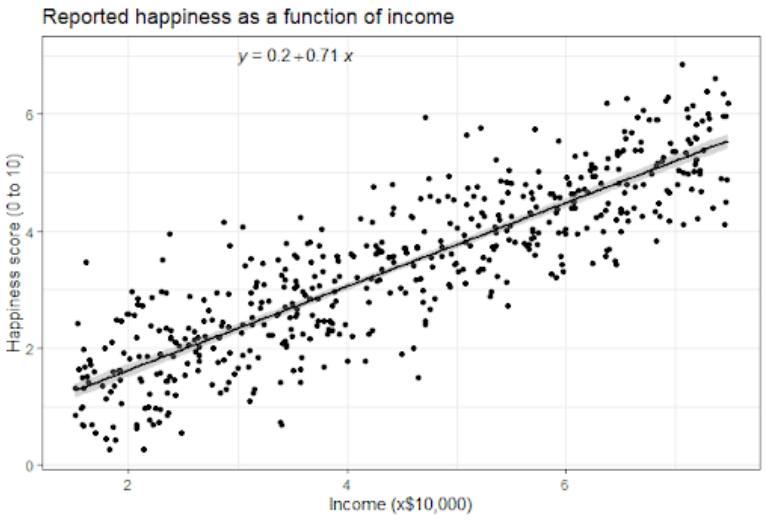


Figure 1: Linear regression of Income on Happiness. Source: Bevans, R. (2020), Simple Linear Regression | An Easy Introduction & Examples, Scribr, Statistics. Retrieved from: https://www.scribbr.com/statistics/simple-linear-regression/ Printed with permission.

variables can take all possible values on a continuum: natural numbers or even decimal places. In addition, scale variables have an explicit unit of measurements: number of residents, square kilometres, and currency (e.g. €, £, $ for the examples in **Table 2**). Due to their units of measurement, values of scale variables have an unambiguous order and distances between these values can be computed precisely.

4. Causality

Regression analysis is a type of statistical analysis that explores causal quantitative relationships between variables. In statistics variables in a causal relationship are habitually expressed by X and Y: X is the independent (cause) and Y the dependent variable (effect). In the conceptual model of research projects, the causality is presented by an arrow from X to Y. A conceptual model is a scheme that presents the set of supposed (!) causal relations between all variables that are selected by the researcher. Regression analysis, then, will test these relations: do they occur, how strong are they and are they positive or negative? It is recommended in quantitative research of causal relations to sketch a conceptual model of these relation. **Figure 2** shows the conceptual model that can be tested by a basic regression model of one dependent and independent variable (section 6). Figures 3 and 4 contain more independent variables: they represent conceptual models that could be but are not included in the papers by Lu et al. (2019) and Li et al. (2015). These two papers are used to illustrate two more complex regression models that the basic one (sections 7 and 8).

Causality means that changes in the values of X result in a systematic increase or decrease of the values of Y. Imagine level of education and annual income: in the causal relation between these two variables level of education is X and annual income is Y, not inversely. In the case of four-digit postal code areas in inner cities as a unit of research, a causal relationship might be found between the proportion of total length of streets that are Pedestrian Only (X) and the Degree of Liveliness (Y). As section seven shows, this is only possible when the broad concept Degree of Liveliness is first operationalised by a single measurable quantitative indicator.

Figure 2: Conceptual model of basic regression model

Individuals with the same level of education rarely have exactly the same annual income. But to make regression analysis worthwhile, a certain trend of systematic association between the values on both variables of all cases has to be visible in a scatterplot. Each dot in the scatterplot in **Figure 1** presents the score of one case (an individual person) in the sample on two variables (Level of Education and Annual Income). Regression analysis estimates both the strength and the direction of causal relationships between variables. The more narrow the point cloud in a scatterplot, the stronger the relationship between X and Y. The direction of the cloud indicates if the relationship is either positive (upward from left to right) or negative (downward). The relation between education and income in **Figure 1** is positive: an increase in educational level causes an increase of income.

5. Inferential statistics

Sometimes ready-to-use databases can be obtained from (semi-)public institutions. If that is not available for your project, you unfortunately need to gather data yourself. That is mostly done by means of a self-prepared survey (questionnaire). Most often the survey has to be conducted with a sample of respondents from the population by means of a well-considered sampling procedure. A sample is inevitable either when the population is too large to be included entirely in the survey or when the population is unknown, i.e. when you can't know exactly who does and who does not belong to it prior to the questionnaire. The entire adult population of a city is too large for a survey and the population of visitors of an urban tourist bubble on a predefined day in the holiday season is unknown in advance.

Regression analysis is an inferential statistical technique: it infers quantitative properties of the entire population from the data obtained with the sample. In statistical language, such an inference is called an estimation. A sample of, say, 500 adults from the civil registry of Amsterdam already yields pretty accurate estimations of the strength and direction of causal relations between variables in the entire population of the city.

6. Basic regression model

The basics of regression analysis can be best explained by the *binary linear model* (equation 1). This model contains one dependent (Y) and one independent variable (X). Both are scale variables and the relation between them is assumed to be linear, meaning that the regression function that defines the model is a straight line. The key numerical parameters of the model are the constant or *intercept* (a) and the *regression coefficient* (b). The third parameter, the *prediction error* (e), is key in the mathematical process of estimation of a and b but can be ignored here.

Y = a + bX + e

(Equation 1)

Take for the example the relation between the two scale variables level of education (X) and annual income (Y) of **Figure 1.** The unit of scale of X is the number of completed years of education starting

with the first year in secondary level and that of Y is €10,000/year. **Figure 2** shows the conceptual model of this relation.

The parameters of the regression line in Figure 1 are estimated by statistical software, such as SPSS: a = 0.2 and b = 0.71 (equation 2).

Y = 0.2 + 0.71X

(Equation 2)

The regression coefficient b predicts the increase of Y if X increases by one unit on its value scale, i.e. one more year of secondary or tertiary education causes an increase of income by € 7,100/year. The intercept a is the value of Y for X = 0: a person with less than secondary education has an estimated income of only € 2,000/year, reflecting that it is most probably earned by an unskilled job. As said, a regression line can also be defined by a negative value of b. In that case, increases in X causes decreases in Y.

7. Multiple linear model

The multiple linear regression model (equation 3) is an extension of the basic model with additional independent variables. A regression analysis includes one and only one dependent variable Y, but the number of independent variables (X_1 to X_n) is limited for practical rather than theoretical reasons. In the multiple linear model, Y is also a scale variable. The independent variables X_1 to X_n are often scale variables but can also be ordinal or nominal.

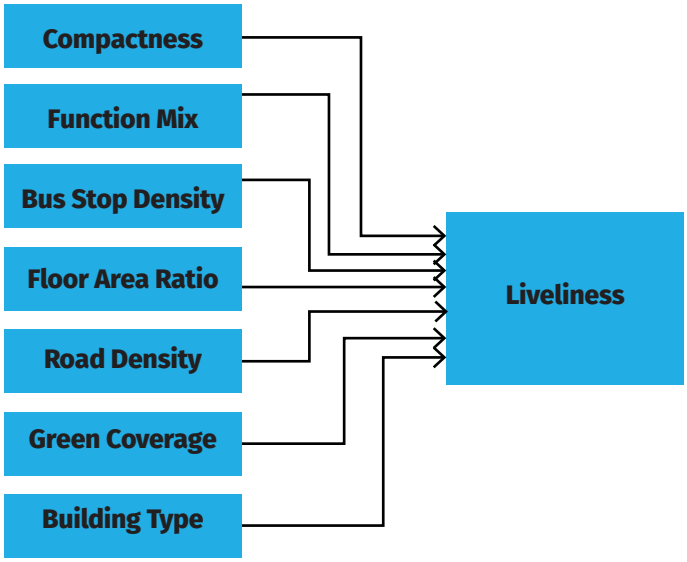


Figure 3: Conceptual model for Lu et al. (2019).

$$Y = a + b_1X_1 + b_2X_2 + ... + b_nX_n$$

(Equation 3)

Multiple linear regression is explained here by an edited version of the output of a study by Lu et al. (2019). The study analyses the impact of seven features of the built environment (X_1 to X_7; **Figure 3**) of the inner city of Beijing on its liveliness (Y).

The cases are 113 RPMUs (Regulatory Planning Management Units); small areas in Beijing's inner city. Liveliness is a multidimensional concept that includes, for instance, available amenities, numbers of people out on the streets and outdoor activities. Because there can be only one Y in regression analysis, the authors choose the scale variable Number of Check-ins on the micro-blog *Sina*, a major social media platform in China. This variable indicates human behaviour. The locational data of check-ins that is required to know in which RPMU people exactly check in is accurate to a single meter. The sample of check-in data covered the first week of

Independent variables	Scale of measurement	Standardized regression coefficients
Compactness index	ratio	-2.09
Function mix index	ratio	1.25*
Bus stop density index	ratio	1.63*
Floor area ratio	ratio	0.46*
Road density index	ratio	0.05
Green coverage index	ratio	0.03*
Building type	(nominal)	
office towers	dummy	-0,95*
modern shopping	dummy	1,25*
Adjusted R^2		0.52

* $p < 0.10$

Table 3. Output of linear regression of built environment features on urban vitality, Beijing. Based on Lu et al. (2019), authors' adjustments.

September 2016 and amounts to a total of 124,658 recorded check-ins spread across the 113 RPMUs.

Table 3 presents the key data of the regression analysis: regression coefficients of the independent variables, the level of statistical significance of each independent variable (the asterisks with its coefficient) and the 'explanatory power' (Adjusted R^2) of the seven independent variables together. The intercept is not included in the Table. It depends on the research question how relevant the intercept is, but mostly it tells nothing really relevant.

Each regression coefficient estimates the impact of the corresponding independent variable on the number of check-ins while controlling for all other independent variables. Controlling means holding these other variables constant. Note that the regression coefficients in **Table 3** are standardised. Standardisation is a mathematical operation that makes the magnitude of impacts of all X on Y comparable, i.e. independent of their distinct value scales. They allow to conclude that the compactness index has the strongest negative (-2.09) and bus stop density the strongest positive (1.63) impact on the number of check-ins. Furthermore, the impacts of road density and green coverage are positive but very limited in strength.

Crucial for the interpretation of a regression coefficient is the asterisk with it. **Table 3** shows that the coefficient of each independent variable has one, except compactness index and road density index. An asterisk shows that the regression coefficient is *statistically significant*. Statistical software estimates a value, the *p-value*, for each independent variable that indicates the probability that the impact of that variable on Y occurs not by a true effect in the population but 'by chance'. 'By chance' is possible because the estimation of the coefficient is based on only a sample of cases from the population. A lower p-value means a higher probability of a true effect in the population. The researcher her- or himself decides on the maximum *p-value* that (s)he accepts. These are commonly either 0.10, 0.05 or 0.01, meaning that (s)he can be respectively 90%, 95%, or 99% sure that the independent variable in question has a true effect on Y in the population. Lu et al. (2019) decided to put that threshold at the 0.10 level ($p < 0.10$). Because inferential statistics infers quantitative properties of the population on the basis of a sample, this threshold cannot be as low as $p = 0.00$,

i.e. 100% certainty.

Adjusted R^2 indicates the explanatory power of the regression model. In statistical terms, the higher adjusted R^2 the higher the proportion of the variance of Y, a statistical measure for variation of it, that is explained by the set of independent variables. The value 0.52 in **Table 3** means that 52% of the variance of the number of check-ins across the RPMUs is explained by the seven selected independent variables. This may seem a disappointingly low proportion but the explanation of more than half of its variance by only seven features of the RPMUs built environment is in fact a good result! The value of the adjusted R^2 increases with the addition of more explaining independent variables. The value 1.00 means that all explaining independent variables are included in the model: an ideal but highly unlikely situation.

Independent variables in multiple linear regression models can be nominal and ordinal. To estimate their causal effects on Y, these have to be transformed into dummy variables or *dummies*. A dummy is a categorical variable with the values 0 and 1. A nominal or ordinal variable with n categories (values) is converted into (n-1) separate dummies. The remaining category is the *reference category*. In **Table 3**, the nominal variable 'building type' indicates the dominant building type in an RPMU. It has three categories: traditional residential buildings, office towers, and modern shopping streets and mall. If we define traditional residential buildings as a reference category, the two (3-1=2) dummies estimate the effect on the number of check-ins of, respectively, office towers and modern shopping spaces as dominant building types relative to that of the reference category. With 1,000 check-in records on *Sina* as unit of scale, the regression coefficient -0.95 of office towers predicts 950 records less in RPMUs dominated by office towers than in RPMUs where old residential buildings are dominant. On the other hand, the regression coefficient 1.25 predicts 1250 records more in RPMUs dominated by modern shopping spaces. Hence, RPMUs dominated by office towers are less popular and RPMUs dominated by modern shopping spaces are more popular to visit than those dominated by traditional residential buildings.

8. Logistic regression model

In the logistic regression model, the dependent variable Y is a categorical one, usually nominal. In case it has two values, logistic regression is binary and if it has more than two values it is *multinomial*. This section is about the binary one. Equation 4 shows its basic model. The two values of the nominal dependent variable are coded 0 and 1.

$$\ln (p_1/p_0) = a + bX + e$$

(Equation 4)

The term $\ln (p_1/p_0)$ that serves as the dependent variable is a *logit*: i.e. the natural logarithm (ln) of a probability ratio. In Equation 4 this is the ratio of the probability that a case in the sample scores the value 1 of the dependent variable (p_1) divided by the probability that it scores 0. A probability is a number between 0 and 1 (Equation 5) and p_0 and p_1 are mutual exclusive (Equation 6).

Independent variables Presence in 500 m radius around hotels	Type	B	Exp(B)
Commercial Floor Space (x 1000 m^2)	ratio	.98**	2.66
Number of Metro Stations	ratio	-.02	0.98
Land Use Mix Index	ratio	.54*	1.72
Number of Cultural Attractions	ratio	-.15	.86
Number of Shopping Attractions	ratio	.85**	2.33
Topography			
False flat	dummy	-.21	.81
Hilly	dummy	-.95	.39
Steep gradients	dummy	-2.18**	.11
Nagelkerke Pseudo R^2		.232	.269

** significant at 0.05
* significant at 0.10
Source: based on Li et al. (2015) authors' adjustments

Table 4: Output of logistic regression of features of built environment on hotel location, Hong Kong. Source: based on Li et al. (2015) authors' adjustments

$$0 > p_0 > 1$$

(Equation 5)

$$p_0 = (1 - p_1)$$

(Equation 6)

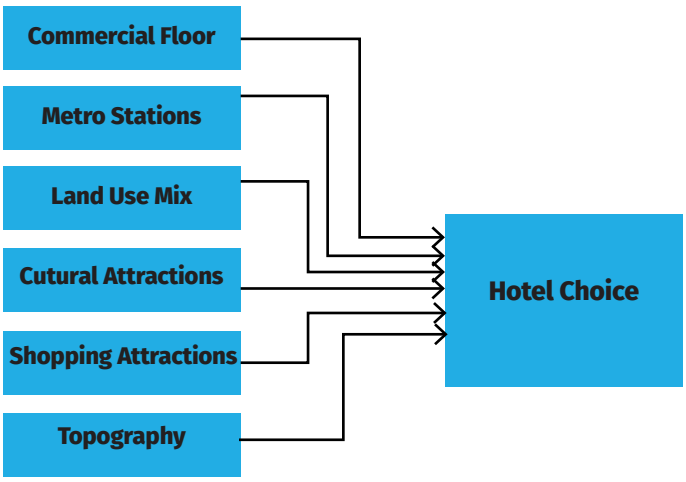


Figure 4: Conceptual model for Li et al. (2015)

Like the multiple linear regression model, the binary logistic regression model can – and usually does – include more than one independent variable (Equation 7).

$$\ln (p_1/p_0) = a + b_1X_1 + b_2X_2 + ... + b_nX_n + e$$

(Equation 7)

Equation 7 shows that the multiple logistic regression model is also additive: contributions of the independent variables are simply added up to predict the value of the logit.

The case study by Li et al. (2015) uses a logistic regression model to analyse how the spatial distribution of types of hotels in Hong Kong are explained by a number of features of their urban environment. **Table 4** is an edited and simplified version of the output of that analysis. The dependent variable Hotel is reduced to an ordinal one with two values: upper- and lower-grade. These grades differ in service levels and room rates. The independent variables are six different qualities of hotels' surrounding urban environments within a radius of 500m. Five of these are scale variables – either absolute numbers or indices. The sixth, topography, is an ordinal variable with four categories: flat land, false flat, hilly, and steep gradients. It is split into three dummies with flat land as the reference category. The conceptual model of this analysis **(Figure 4)** has the same structure as that of the multiple linear model.

The overall objective of the analysis is to examine if the existence of different types of clustered tourist districts can be conceptualised. It is assumed that the two different types of hotels are

visited by types of guests, i.e. tourists with different preferences of service levels set against room rates. Starting from that assumption, the logistic regression analyses if these two different types of hotels are surrounded by built environments with different values of these six qualities, i.e. values that fit better with these different types of guests' demands and budgets.

The regression coefficients B in **Table 4** express the impact of the six independent variables on the logit ln ($p_{upper\ grade}$ / $p_{lower\ grade}$) of hotel choice. It is practically impossible to know what a change in this natural logarithm means for changes in p1 and p0. To get rid of the natural logarithm, Exp (B) expresses the change in just the probability ratio ($p_{upper\ grade}$ / $p_{lower\ grade}$). Every additional amount of 1,000m2 of commercial floor space closely around a hotel increases the probability that it is an upper-grade hotel divided by the probability that it is a lower-grade one by a factor 2.66. The number of shopping attractions has a highly similar effect (2.33) on this probably ratio. The effects of both variables are statistically significant at the 95% *level of confidence*. In reverse, the probability that a hotel that is surrounded by steep streets, compared to flat land as the reference category, is an upper-grade one is nine times as low (0.11) as the probability that it is a lower-grade one. This indicates that lower budget guests accept the inconvenience of steep streets much more easily to save on the room rate in a hotel.

In linear regression models, adjusted R^2 indicates the amount of explained variation of the continuous variable Y by X_1 - X_n. For logistic regression with a categorical dependent variable, a few pseudo R^2 measures are available. Pseudo indicates a lower level of precision than of adjusted R^2 in linear regression. The advantage of the nagelkerke pseudo R^2 that is used by Li et al. (2015) is the range of its values between 0 and 1, just like adjusted R^2. The value of the nagelkerke pseudo R^2 (.232) is moderate, demanding for some more independent variables to explain hotel choice.

9. Discrete choice models

The types of regression analysis presented so far explain revealed human behaviour as being triggered by existing spatial qualities of the built environment in a specific location. That knowledge can be useful to evaluate the appreciation of qualities by users of the built environment. Discrete choice analysis that is presented in this last section analyses stated, i.e. planned intentional behaviour. Discrete choice models are based on multinomial logistic regression (MNL). Discrete choice analysis is largely unknown in urban design and spatial planning: a great pity because it can hit on the preferences of (potential) users for desired future spatial qualities of project locations. These qualities can already be existing in that location, or maybe not. The opportunity to include knowledge of users' appreciation of not yet present spatial qualities adds an important dimension to the utility of regression analysis for an urban design or planning process.

One of the rare examples of the use of discrete choice analysis in urban design is a project by Susanne van Rijn (van Rijn, 2020) in the municipality of Westland, the Netherlands. The objective of the project is to identify the appreciation of spatial qualities of outdoor public space by adolescents in the age range twelve to seventeen to take exercise, i.e. to become healthier.

Based on an extended literature review, van Rijn

No.	Attributes	Explanation	Level
1	Vegetation	Amount and variation	0 = little to no vegetation; little variation 1 = much vegetation and variation in the public space
2	Barriers	Physical barriers that hamper accessibility	0 = broad busy traffic roads often causing waiting times 1 = only quiet street rarely causing waiting times
3	Facilities	Facilities for sports and play	0 = none or few 1 = many, diversity of types
4	Paths	For cycling and walking	0 = only around public space 1 = around and through public space
5	Proximity	Walking distance from home	0 = more than 5-minute walk 1 = at most 5-minute walk
6	Lighting	Quality of lighting	0 = large parts of public space not illuminated 1 = public space is sufficiently illuminated
7	Seclusion	Spots where one is invisible from surroundings	0 = present 1 = absent
8	Water	Water features	0 = absent 1 = present
9	Seating	Open air furniture to sit	0 = absent 1 = present
10	Toilets	Public toilets	0 = absent 1 = present

Table 5: Selected attributes for physical activity in public space. Source: based on Van Rijn (2020), author's adjustments.

first carefully selected ten key attributes of public spaces that are supposed to influence physical activity behaviour of adolescents **(Table 5)**. Attributes in discrete choice analysis are equivalent to categorical variables in the above discussed regression models. The categories (values) of attributes are called 'levels'.

Each attribute in **Table 5** has two levels: 0 and 1. One level is assumed to be positively associated with taking exercise and the opposite holds for the other level. Attributes in discrete choice analysis can have more than two levels, but for reasons of validity and interpretability these are rarely more than three.

Next, sixteen different alternatives were composed: A to P in **Table 6**. Alternatives are imaginary constructs: deliberately composed combinations of levels of the ten selected attributes. By means of an online questionnaire with the software Qualtrics, a sample of adolescents was asked to make choices between alternatives as they would do in the real world, i.e. in case these alternatives would really exist. Each respondent answered five questions. In each question two alternatives out of the sixteen (A to P) were randomly combined and the respondent was asked which of these two (s)he prefers to take exercise in. The option 'neither of the two' was also possible. To enable them to choose, the alternatives were made visible with drawings, including a brief explaining text to emphasise some features of the alternatives (see Figures 5 and 6 for examples). These are in fact simple spatial designs for the project location. The questionnaire yielded a dataset of 309 valid cases (N in **Table 7**).

Sixteen alternatives is a very low number if one realises that the total number of different alternatives in case of ten attributes with two levels each equals 1,024 (2^{10}). In general, over one thousand

	Attributes	Alternatives															
		A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	Vegetation	0	0	0	0	1	1	1	1	1	1	1	1	0	0	0	0
2	Barriers	0	0	1	1	0	0	1	1	0	0	1	1	0	0	1	1
3	Facilities	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1
4	Paths	0	1	1	0	0	1	1	0	0	1	1	0	0	1	1	0
5	Proximity	0	0	1	1	0	0	1	1	1	1	0	0	1	1	0	0
6	Lighting	0	1	0	1	1	0	1	0	0	1	0	1	1	0	1	0
7	Seclusion	0	1	1	0	1	0	0	1	1	0	0	1	0	1	1	0
8	Water	0	1	1	0	0	1	1	0	1	0	0	1	1	0	0	1
9	Seating	0	0	1	1	1	1	0	0	0	0	1	1	1	1	0	0
10	Toilets	0	1	0	1	1	0	1	0	1	0	1	0	0	1	0	1

Table 6: Designs of imaginary public spaces. Source: based on Van Rijn (2020).

alternatives are far too many in the practice of a project. Substantial reduction of the number of alternatives, without losing the possibility to estimate parameters with MNL, is made possible by using the appropriate basic plan that matches this 2^{10} situation as developed by the American mathematician Sydney Addelman in the 1960s (Steenkamp, 1985).

As Equations 8A to 8P show, the MNL model in discrete choice analysis is not one single equation, but one for each of the sixteen alternatives in the study

$$V_A = ASC_A + \beta_1 X_{A1} + \beta_2 X_{A2} + ... + \beta_{10} XA_{10}$$

(Equation 8A)

$$V_B = ASC_B + \beta_1 X_{B1} + \beta_2 X_{B2} + ... + \beta_{10} X_{B10}$$

(Equation 8B)

-

$$V_P = ASC_P + \beta_1 X_{P1} + \beta_2 X_{P2} + ... + \beta_{10} X_{P10}$$

(Equation 8P)

Figure 5: Alternative O with the highest calculated utility. Source: Van Rijn (2020, 80)

Figure 6: Alternative J with the lowest calculated utility. Source: Van Rijn (2020, 80)

Attributes	Level of reference	β	Statistical significance
Vegetation	little to no vegetation; little variation	-0.403	0.000000846*
Facilities	none or few	-0.368	0.0000142*
Barriers	only quiet street rarely causing waiting times	0.255	0.00275*
Proximity	home is further away than 5 minute walk	-0.169	0.0458
Lighting	public space is sufficiently illuminated	0.152	0.0733
Water	water feature present	-0.125	0.123
Seclusion	visible throughout entire surroundings	0.097	0.239
Paths	only around public space	0.062	0.463
Seating	seating furniture absent	0.049	0.556
Toilet	public toilets present	-0,015	0.871
Sample	309		
Rho squared	0.223		

Table 7: Estimation of attribute parameters. Source: van Rijn (2020: 78).

The values of X_1 to X_{10} equal the levels of its 10 attributes in each alternative as defined in **Table 6**. The MNL analysis that was carried out with the software package Pandas Biogeme estimated two types of parameters: ten regression coefficients (β1 to β10) over all sixteen alternatives and one alternative specific constant (ASCA to ASCP) for each alternative. Together, these estimated and predefined values can be used to calculate the total utility of each alternative (V_A to V_P: **Table 8**).

Table 7 shows the output of MNL. The values of the βs represent the relative contribution of each attribute to the appreciation of public space for physical activity by adolescents. In absolute value, the relative contribution is highest for vegetation (-.403). Its negative sign indicates that the reference level as defined in Table 7, i.e. abundant and highly varied vegetation, contributes negatively to the appreciation of outdoor public spaces by adolescents 12-17 years of age to take exercise. The sign of β for the attribute 'toilet' shows that the presence of public toilets in public space is not appreciated positively by the adolescents to go there to take exercise. However, its very low absolute value (0.015) indicates that the weight attached to their absence is in fact very limited. Moreover, its p-value (.871) is much larger than 0.10, showing that it is not 90% sure that 'toilet' has any effect at all on adolescents' appreciation of public space for exercise. The same holds for the attributes 'water', 'seclusion', 'paths', and 'seating'. It is important, finally, to realise that the β values only give relative comparisons of the weight of attributes: they are categorical and lack a unit of measurement (section 3).

Table 8 shows the calculated total utilities for the alternatives A to P that are defined by the 2^{10} basic plan. Figures 5 and 6 show, as examples, the alternatives with the highest (O) and the lowest (J) total utility. You may think that full implementation of the alternative with the highest utility is the basis for the best possible spatial design or plan in the real world. That is, however, not necessarily the case. Because the basic plan that defines alternatives is based on mathematics and has no empirical connection to any urban design or planning context, the one with the highest total utility can include attribute levels that contribute negatively. Moreover, it is possible that local conditions make it impossible to realise a specific attribute level, despite how highly that might be appreciated. Imagine that the absence of Barriers appears important but the site is located at a major road which cannot be altered. In

Alternative	Utility function	Utility value*
O	V_O	3.13
C	V_C	2.30
A	V_A	1.94
D	V_D	1.88
K	V_K	1.63
E	V_E	1.62
M	V_M	1.47
B	V_B	1.40
G	V_G	1.36
P	V_P	1.26
L	V_L	1.13
N	V_N	1.06
I	V_I	0.69
H	V_H	0.50
F	V_F	0.34
J	V_J	0.31
0	V_0	0

*: rounded to two decimals

Table 8: Estimated utility values of alternatives. Source: van Rijn (2020: 78)

fact, the β-values give more specific information of the attributes and are therefore often more valuable for use in urban design or planning.

To conclude, the results of the MNL show the relative importance of selected attributes for use in an urban design as expressed by their β-values. But that still does not mean that 'the one and only' design follows straightforwardly. In fact, it still says little about how that design should be realised. The way in which attributes are combined into a composition of space in the real world is where the expertise of urban designers like you play a key role.

10. Conclusion

Master's graduation projects like yours would result in well-elaborated and highly integrated proposals for spatial qualities in urban designs or spatial policies for your project location. If so, in your case it is highly likely that you would want to base your proposal for spatial qualities on users' revealed or stated spatial behaviour. Regression analysis accurately estimates the quantitative amount of contribution of each separate feature of spatial quality to the explanation or forecasting of users' spatial behaviour. The examples of the use of regression models that are discussed in the sections seven, eight, and nine explain revealed spatial behaviour (visiting specific places in central Beijing and types of hotels in Hong Kong) or stated behaviour (physical activity in Westland) by selected spatial features of these locations' built environments. This type of knowledge can be significant for appropriate urban design or spatial policies in your project as well.

Not included in the examples are personal socio-demographic characteristics, like age, gender, or educational level, as independent variables. Including these characteristics might show that the relationships between built environment and behaviour is quantitatively different for different subgroups. That can be done in two ways. First, such characteristic can be included as additional independent variables in the regression model or the discrete choice model. Another way is to split the sample of users into subgroups according to such characteristics and run a regression analysis for each subgroup.

It should be noted that the use of regression analysis as a quantitative method is not mutually

exclusive from qualitative research methods: it is not a matter of choosing one or the other in your urban design or planning project. On the contrary, if you consider using regression analysis it is still essential to first construct an adequate conceptual model and then think very carefully about which variables in that model should be inserted into your regression model. Hence, a thorough review of relevant international literature on our research topic is required. Overall, qualitative methods like the review and also analysis of policy documents or interviews with key persons are crucial (!) in all stages of the iterative cycle of research and design during the project.

In a 2013 paper by Emeritus Professor of the Faculty of Architecture and the Built Environment, Taeke de Jong, commented that 'the specialised probabilities or even "truths" of empirical research [...] cannot be successfully integrated in [a design of] one spatially, ecologically, technically, economic, cultural and managerial unique case' (de Jong, 2013: 22). Key features of regression analyses are unidirectional causal reasoning, inclusion of only a limited number variables, and single moment bound data collection. Hence, it is not a panacea in its own right in dealing with the complexity of interwoven multidimensional challenges of designing unique locations. But the fact that it 'cannot be successfully integrated' underestimates the usefulness of empirical research techniques like regression analysis in urban design and planning projects. Just like yours!

11. References

De Jong, T. (2013). Empirical research and spatial design. *Atlantis*, 23(3), 22-25.

Li, M., Fang, L., Huang, X., & Goh, C. (2015). A spatial–temporal analysis of hotels in urban tourism destination. *International Journal of Hospitality Management,* 45, 34–43.

Lu, S., Shi, C., & Yang, X. (2019). Impacts of built environment on urban vitality: Regression analyses of Beijing and Chengdu, China. *International Journal of Environmental Research & Public Health*, 16, 1-16.

Steenkamp, J.-B. (1985). De constructie van profielensets voor het schatten van hoofdeffecten en interacties bij conjunct meten. In *Jaarboek voor de Nederlandse Vereniging van Marktonderzoekers* (pp. 125-154). Nederlandse Vereniging van Marktonderzoekers. https://www.researchgate.net/publication/40161511_De_constructie_van_profielensets_voor_het_schatten_van_hoofdeffecten_en_interacties_bij_conjunct_meten

Van Rijn, S.E.M. (2020). *Urban design for physical activity. An exploration of the role of the urban design of public spaces in stimulating adolescents to be more physically active in Westland, the Netherlands* (master's thesis). TU Delft. http://resolver.tudelft.nl/uuid:5eaed236-732e-405a-a1d9-930059ca8224

12. Further reading

Allison, P.D. (1999). *Multiple Regression. A primer.* Pine Forge Press.

Romein, A., & Maat, K. (2013). Spatial organisation of consumer services in the polycentric urban context: A travel behaviour approach of cinema-going in the city-region of Rotterdam. *Tijdschrift voor Economische en Sociale Geografie*, 104(4), 491-509.

Mambretti, I.M. (2007). *Urban Parks between Safety and Aesthetics - Exploring urban green space using Visualisation and Conjoint Analysis methods* (doctoral thesis). ETH Zurich.