International Journal of Export Marketing

ISSN online: 2059-0903 - ISSN print: 2059-089X
https://www.inderscience.com/ijexportm

Spatial proximity as an independent variable in (international) marketing and management research

André Van Hoorn

DOI: 10.1504/IJEXPORTM.2022.10051730

Article History:

Received:	17 January 2022
Accepted:	03 October 2022
Published online:	02 February 2023

Spatial proximity as an independent variable in (international) marketing and management research

André Van Hoorn

Radboud University,
Institute for Management Research,
P.O. Box 9108, 6500 HK,
Nijmegen, The Netherlands
Email: Andre.vanHoorn@ru.nl

Abstract: Spatial proximity matters for a range of managerial and marketing-related phenomena and its use as an independent variable seems innocuous. However, building on recent advances in the contextual distance literature in international management, we argue that analyses involving spatial proximity can be far more intricate than commonly realised. The problem is the statistical biases that can be caused by location-identification, which refers to proximity measures' tendency to correlate with locations in a sample and thus partially capture location-specific effects when included in regression models. Empirical evidence confirms that spatial proximity measures are also prone to location-identification. Further evidence demonstrates how location-identification matters for spatial proximity research because it can bias the estimated relationship between spatial proximity and the phenomenon studied. Our main contribution is to show the relevance of location-identification threatening the validity of spatial proximity studies that are common in international marketing and in management studies more broadly. Export managers are also advised to be wary of conflating distance-driven influences on their firms with effects genuinely due to location factors.

Keywords: location effects; spatial proximity effects; distance-location conflation; distance-location correlations; pure-distance approach; fixed-effects approach.

Reference to this paper should be made as follows: Van Hoorn, A. (2022) 'Spatial proximity as an independent variable in (international) marketing and management research', *Int. J. Export Marketing*, Vol. 5, No. 2, pp.163–182.

Biographical notes: André van Hoorn is a Professor of International Economics and Trade in the Department of Economics and Business Economics at Radboud University, the Netherlands. His research interests concern the implications of differences and distances between business environments for economic Actors such as multinational enterprises that are embedded in multiple business environments simultaneously. André has published pieces in journals from different disciplines, including *the Journal of International Business Studies*, *The Journal of Development Economics*, *The Journal of Cross-Cultural Psychology*, and *Global Strategy Journal*.

1 Introduction

Researchers in international marketing and management commonly estimate regression models in which variation in some variable is explained by variation in a measure of spatial proximity. This paper considers confounding variables that can correlate with distance and undermine the validity of estimated effects of spatial proximity. Spatial proximity has been linked to a range of phenomena in marketing and management (e.g., Cannon and Homburg, 2001; Funk, 2014; Knoben and Oerlemans, 2006; Ragozzino and Reuer, 2011; Todri et al., 2022). However, for international marketing and management, spatial proximity is quintessential (Ghemawat, 2016). Spatial proximity refers to how the locations of two units are placed relative to each other. In empirical research, a variety of measures is used for operationalising the concept of spatial proximity. However, most studies in marketing and other management fields consider a measure of geographic distance.[1]

Empirically, studies of how spatial proximity matters for international marketing and management involve estimating a regression model with geographic distance between pairs of locations (e.g., home- and host country dyads) as the explanatory variable. This approach is standard and seems innocuous. Nevertheless, recent methodological critique of empirical studies of the effects of cross-national cultural and institutional distance give reason to pause and consider the use of spatial proximity as an independent variable as well. In a nutshell, the critique is that measured cultural and institutional distance correlate strongly with home and host countries' cultural and institutional profiles, which biases estimated effects of these contextual distances (e.g., Brouthers et al., 2016; Van Hoorn and Maseland, 2014, 2016). The underlying mechanism is that cultural and institutional distance are measured using countries' cultural and institutional profiles, which is problematic when the sample considered comprises only one home country or only one host country. In such a scenario, there will automatically be a strong correlation between the two types of constructs, countries' cultural or institutional profiles on the one hand and cultural or institutional distance between countries on the other. By construction, the above critique is distinctive to contextual distance measures involving so-called difference scores for countries' contexts for doing business, particularly cultural and institutional distance (Beugelsdijk et al., 2017; Brouthers et al, 2016). However, a recent extension of this critique by Van Hoorn (2020) could imply that a similar problem occurs when considering spatial proximity as well. The issue raised by Van Hoorn (2020) is that contextual distance metrics may not only correlate with observable country-specific contextual factors such as national culture or institutions but also with country or location itself. Distance indicators are argued to have the ability to identify specific countries in a sample, what Van Hoorn (2020) refers to as 'location-identification.' Hence, cross-national distance indicators can partly capture the confounding influence of any country-specific factor when used as an independent variable.

If geographic distance is also prone to location-identification, this can have important implications for empirical spatial proximity research. As an example, consider a designer clothing chain with its head office located in downtown São Paulo that owns several stores also located in downtown São Paulo. The chain has recently opened a branch in Buenos Aires, but this branch is performing poorly. Senior management attributes this poor performance to the difficulties of managing an establishment that is not close to the head office. If senior management is correct, this would be a genuine effect of spatial proximity. However, it could also be that this poor performance is due to demand for designer clothing in Buenos Aires being lower than in downtown São Paulo. If so, this would be a location-specific effect that is distinct from and should not be conflated with spatial proximity effects. In particular, the two types of challenges, distance-driven vs. location-driven, would require different managerial responses. Distance-driven challenges, for example, may involve communication difficulties that could be mitigated through revised communication protocols and additional ICT support. These solutions, however, do nothing to address a lack of local demand for designer clothing. When a measure of spatial proximity partly captures location-specific influences, we can never be sure what effect or effects are reflected in the estimated coefficient for this spatial proximity measure.

This paper seeks to address two related issues. The first is whether, in addition to measures of cross-national contextual distance, measures of spatial proximity also correlate with country (i.e., whether these measures are also prone to location-identification). Following our theoretical and empirical evidence on this first issue, the second issue that we take up concerns the occurrence of biases in the estimated effect of spatial proximity because of such location-identification.

In the next section we continue with an explanation of the theoretical mechanism of how measures of spatial proximity might be able to provide partial identification of one or more locations in a sample. In this section we also analyse large, publicly available datasets to present evidence on the degree to which measured spatial proximity, particularly geographic distance, indeed correlates with and partially identifies locations in real-world samples. In the third section, we discuss and illustrate empirically how location-identification can bias the estimates for regression models that include geographic distance as an independent variable. We end with a brief discussion and concluding remarks concerning the future use of spatial proximity as an independent variable, not just in international marketing and international management but in all fields of management research for which spatial proximity is relevant. Section 2 includes a conceptual and rather abstract explanation of the mechanism driving location-identification. Section 3.1 similarly provides a technical discussion of the potential for biases in estimated distance effects. Readers that are less interested in such details may skip one of or both these sections. Instead, they can only gauge the empirical results presented in Tables 1–4 or move forward to Section 3.2. They may even skip Sections 2 and 3 altogether, moving straight to the discussion and recommendations that we present in Sections 4 and 5.

2 Spatial proximity and partial identification of country locations in samples

2.1 Theoretical background

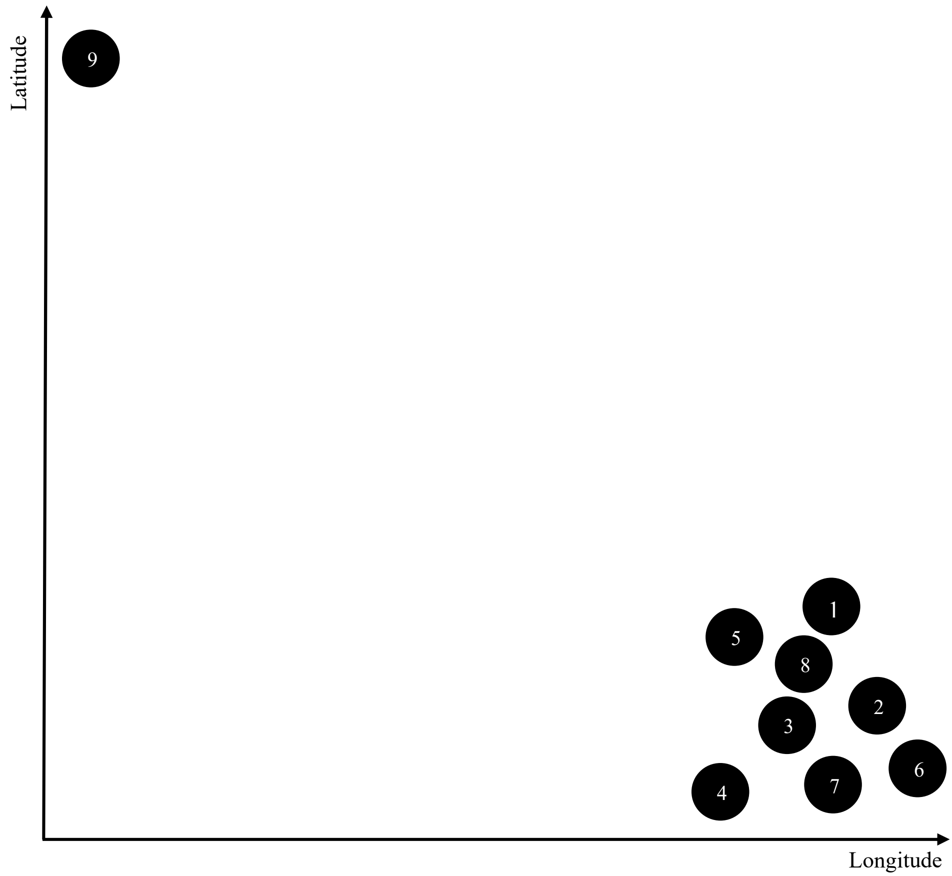
How might measures of spatial proximity or geographic distance be able to identify locations in a sample? Van Hoorn (2020) makes the claim that cross-national distance indicators partially identify locations in the context of home-host country difference scores that collapse contextual dissimilarities into a single, one-dimensional contextual distance index. This collapsing is quite common in international marketing and international management but naturally raises concerns about standard floor and ceiling effects (see Cramer and Howitt, 2004 and Everitt, 2002 for textbook discussions). In one-dimensional space, distance measured as a home-host country contextual difference score can have only two directions, 'up' or 'down.' From a fixed point of origin, distance can therefore only go towards two possible locations, one location that is 'below' the point of origin and one location that is 'above' the point of origin. Having to predict the destination of where a given amount of contextual distance will take you, you would thus automatically be correct at least 50% of the time. Moreover, this percentage increases drastically when countries or locations are located at the extremes of the distribution or scale. For such countries, distances, particularly large distances, can often go in only one direction, thus providing almost exact identification of the countries involved (see Panel b of Figure 1 in Van Hoorn, 2020 for a graphical illustration). Geographic distance, in contrast, is fundamentally two-dimensional, implying that it can unfold in an infinite number of directions.[2] Hence, it could be that spatial proximity is much less prone to location-identification.

Nevertheless, digging deeper, we find that Van Hoorn's (2020) argument on how measured cross-national contextual distance can partially identify countries or locations likely applies to two-dimensional geographic distance as well. Underlying Van Hoorn's (2020) argument is the idea that location-identification occurs whenever one or more locations in a sample have such a unique association with measured distance to/from these locations that measured distance partly identifies these locations. This, in turn, happens when locations are not placed perfectly randomly in the relevant space. This space can be a one-dimensional space, which is the setting for Van Hoorn's (2020) analysis. However, a priori there seems little reason to assume that the relevant space must be one-dimensional and cannot be multidimensional. Similarly, there does not seem to be anything in Van Hoorn's (2020) argument that stipulates that the distance measure considered should concern a difference score.

When there is much randomness in the way locations are spread in a given space, there is no systematic relationship between a location and the distance or proximity of this location to the other locations in a sample. When the locations in a sample are not placed purely randomly in space, in contrast, variation in measured proximity will automatically be specific to one or more locations in the sample. It thereby does not necessarily matter whether this non-randomness occurs in one-dimensional space or in two-dimensional space. Instead, the relevant issue is whether there is sufficient non-randomness to ensure that there is a location in the sample that distinguishes itself from the other locations in the sample because the average proximity to/from this location is either much larger or much smaller than the average proximity among the other locations in the sample. Figure 1 graphically illustrates location-specific variation in spatial proximity with locations placed in two-dimensional geographical space, which contrasts with Van Hoorn's (2020, p.3) illustration of non-randomness in one-dimensional space. Although Locations 1-8 are scattered reasonably randomly in geographical space, Location 9 is a clear outlier location, suggesting that in this scenario measured spatial proximity correlates with and is able to identify Location 9 in particular.

In the real world, locations for instance the home and host countries in which multinational firms are active are not necessarily placed randomly in geographical space. Implication is that measured proximity does not only inform us about the actual proximity between sets of locations but also has the ability to partially identify specific locations in a sample. An example is Location 9 in Figure 1. Such location-identification, in turn, means that measured proximity can partly capture location-specific effects, e.g., lack of demand for designer clothing in a location or other such location-specific influences when used as an independent variable in a regression model.

Figure 1 An example of a location (L9) and the uniqueness of its proximity to the other locations in a sample (L1-l8) in a two-dimensional space that is akin to geographical space



2.2 Empirical evidence on location-specific variation in spatial proximity and location-identification

To what extent does measured spatial proximity indeed correlate with locations in real-world samples? The main answer to this question comes from an analysis that assesses the percentage of total variation in measured spatial proximity that is location-specific. This analysis involves estimating the variance explained for a model with measured spatial proximity as the dependent variable and origin and destination location fixed effects as predictor variables. Values of 0% or 100% for variance explained thereby imply no correlation between proximity and location (0%) and a perfect correlation between proximity and location respectively (100%). However, to be complete, we also provide a direct assessment of measured spatial proximity's ability to differentiate between locations in a sample. This latter analysis involves estimating multinomial models with country as the categorical dependent variable and measured spatial proximity as the independent variable. The corresponding statistical test is whether spatial proximity (statistically) significantly predicts these countries. Since proximity is a characteristic of two locations relative to each other, two multinomial models are estimated for each sample considered, one with origin locations and one with destination locations as the dependent variable. However, for the empirical analysis Table 1, we have samples that are balanced and comprise the same origin and destination countries. Hence, multinomial estimation results are the same for both origins and destinations.

Rows 1 and 2 of Table 1 present results and descriptive statistics for two measures of spatial proximity concerning the majority of countries in the world. Inclusion of nearly all countries in the world results in samples of 217 and 227 origin/destination countries and 46,872 and 51,302 country dyads respectively. Results indicate that location-specific variation in spatial proximity is never trivial and that location fixed effects may account for a substantial amount of differences in measured spatial proximity (Column 2). Hence, it seems that measures of spatial proximity correlate significantly with location. Similarly, multinomial results indicate that measures of spatial proximity significantly predict location and are thus quite able to differentiate between locations in a sample (Column 5).

At the same time, there is noticeable heterogeneity between the measures and samples analysed. Partly, this heterogeneity likely reflects the different nature of the two proximity measures considered. However, sample composition also matters, as illustrated by results for geographic distance presented in Rows 3 and 4 of Table 1. The results in Rows 3 and 4 concern geographic distance and therefore compare to results presented in Row 1, except that they pertain to different samples. However, this influence of sample composition does not detract from the main conclusion that we draw from the empirical evidence in Table 1. This conclusion is that location-identification does not only occur when considering the kind of dyadic distance metrics studied mostly by international marketing and international management researchers (contextual difference scores such as cultural or institutional distance; Van Hoorn, 2020) but also when considering 'plain' spatial proximity.

Rows 1 and 2 of Table 1 present results and descriptive statistics for two measures of spatial proximity concerning the majority of countries in the world. Inclusion of nearly all countries in the world results in samples of 217 and 227 origin/destination countries and 46,872 and 51,302 country dyads respectively. Results indicate that location-specific variation in spatial proximity is never trivial and that location fixed effects may account for a substantial amount of differences in measured spatial proximity (Column 2). Hence, it seems that measures of spatial proximity correlate significantly with location. Similarly, multinomial results indicate that measures of spatial proximity significantly predict location and are thus quite able to differentiate between locations in a sample (Column 5).

Table 1 Location-specific variation in spatial proximity and partial identification of locations

	Column 1: Measure of spatial proximity (kms)	*Column 2: % variation in spatial proximity that is location-specific*	*Column 3: Number of origin/ destination locations*	*Column 4: Mean, SD, and sample size*	*Column 5: Statistical fit of multinomial model with origin or destination location as categorical dependent variable and proximity as the independent variable*
1	Geographic distance	32.11% (95% CI: 31.13, 33.09%)	$N_o = N_d = 217$	8,443 (SD = 4,654) [n = 46,872 country dyads (= 217^2–217)]	p = 0.0000 (Pseudo R^2 = 0.0157)
2	Sea distance	20.29% (95% CI: 19.55, 21.04%)	$N_o = N_d = 227$	10,482 (SD = 5,431) [n = 51,302 country dyads (= 227^2–227)]	p = 0.0000 (Pseudo R^2 = 0.0097)
3	Geographic distance, countries on American continent only	67.14% (95% CI: 62.44, 71.84%)	$N_o = N_d = 48$	3,086 (SD = 2,148) [n = 2,256 country dyads (= 48^2–48)]	p = 0.0000 (Pseudo R^2 = 0.0434)
4	Geographic distance, countries on European continent only	53.52% (95% CI: 48.28, 58.77%)	$N_o = N_d = 37$	1,589 (SD=850) [n = 1,332 country dyads (= 37^2 – 37)]	p = 0.0000 (Pseudo R^2 = 0.0385)

Notes: Results for Column 2 are obtained with spatial proximity as the dependent variable and country fixed effects as the independent. The 95%CIs for the percentage location-specific variation in spatial proximity provide a statistical test of the joint significance of these country fixed effects and are obtained using bootstrapping with 10,000 repetitions. The p-values in Column 5 are for a likelihood ratio test that tests the likelihood of obtaining a particular positive or negative estimate for the explanatory variable (distance) even though the actual coefficient is zero. The measure of geographic distance (Rows 1, 3 and 4) refers to a weighted distance measure that takes into account the geographical distribution of the population within each country.

Source: Data are from the CEPII GeoDist database (Mayer and Zignago, 2011). The measure of sea distance (or bilateral maritime distance) (Row 2) comes from the CERDI-seadistance database (Bertoli et al., 2016). Samples exclude cases where ‘origin’ and ‘destination’ locations are the same. A replication package and complete estimation results are available on request.

At the same time, there is noticeable heterogeneity between the measures and samples analysed. Partly, this heterogeneity likely reflects the different nature of the two proximity measures considered. However, sample composition also matters, as illustrated by results for geographic distance presented in Rows 3 and 4 of Table 1. The results in Rows 3 and 4 concern geographic distance and therefore compare to results presented in Row 1, except that they pertain to different samples. However, this influence of sample composition does not detract from the main conclusion that we draw from the empirical evidence in Table 1. This conclusion is that location-identification does not only occur when considering the kind of dyadic distance metrics studied mostly by international marketing and international management researchers (contextual difference scores such as cultural or institutional distance; Van Hoorn, 2020) but also when considering 'plain' spatial proximity.

3 Statistical biases when using spatial proximity as an independent variable

The previous section considered the possible occurrence of location-identification and hence the potential for spatial proximity to capture location-specific effects when used as an independent variable. However, more directly relevant for researchers is the question how location-identification ends up affecting empirical analyses that include spatial proximity as an independent variable. In the first part of this section we present a conceptual answer to this question. In the second part, we present four empirical examples that illustrate how location-identification matters for spatial proximity studies in (international) marketing and management because it can bias the estimates for regression models that include spatial proximity as an independent variable.

3.1 Location-identification, correlated measurement error, and genuine proximity effects

In line with Van Hoorn (2020), we find that the generic answer to the question posed above is that empirical proximity studies may suffer an endogeneity problem (Wooldridge, 2002). Most obviously, proximity's tendency to correlate with location and therefore capture location-specific influences on the dependent variable means that we can never be sure what effect or effects are reflected in the estimated coefficient for a spatial proximity measure. An example is that what may seem like a genuine effect of geographic distance (e.g., on the performance of a branch of a designer clothing chain that is located far away from the chain's head office) is in fact reflecting the influence of a location-specific factor (e.g., a lack of local demand for designer clothing). Because location fixed effects can capture all sorts of location-specific influences, including plain unobservable location-specific measurement error, it becomes rather difficult to ascertain whether found effects of spatial proximity represent genuine proximity effects or the confounding effect of location-specific influences. We may refer to this specific endogeneity (or omitted variables) problem as the location-identification problem.

Formally, an endogeneity problem, which includes problems due to omitted variables, involves so-called correlated measurement error, which is to say that the chief independent variable in a regression model is measured with measurement error that exhibits a certain degree of correlation with the error term of the regression model. Mathematically, we can write:

$$D = \beta_0 + \beta_1 I + e, e \sim N(0,1), \tag{1a}$$

where

$$I = P + \rho \times e, P \sim N(0,1). \tag{1b}$$

In this basic model with an endogenous regressor, D denotes the dependent variable and I denotes the independent variable. However, the independent variable is measured with measurement error that exhibits a certain degree of correlation (ρ) with the error term of the regression model (e). The variable P (for proximity) represents the independent variable but without measurement error, while the term ~N(0,1) indicates that variables follow a standard normal distribution.

In general, statistical biases due to correlated measurement error become more significant, the stronger the correlation ρ between the measurement error in the independent variable and the error term in the regression model. Similarly, the extent of statistical biases in empirical spatial proximity research are a function of:

1 the strength of the correlation between spatial proximity (P) and location or country

2 the strength of the correlation between the dependent variable of interest (D) and location.

The stronger these two correlations are, the more we can expect that estimated coefficients for spatial proximity are biased. Vice versa, if one of these correlations is trivial, biases due to location-identification are also likely to be small. Estimates can thereby suffer an upward or a downward bias relative to the 'true' value. An upward bias, including an overestimation of true variance explained, occurs when the estimated relationship between spatial proximity and the dependent variable is partly driven by location-specific influences and therefore does not only reflect a genuine spatial proximity effect but (also) the effect of location-specific factors. A downward bias, including an underestimation of true variance explained, occurs when the location-specific influences captured by the proximity measure go counter to the genuine spatial proximity effect. In this case, the two types of effects, a location-specific effect and the genuine spatial proximity effect, may (partially) cancel out, thus lowering the estimated effect of spatial proximity on a dependent variable.

3.2 Empirical illustration of location-identification and (Un) biased empirical model estimates

In the second part of this section we give empirical flesh to the above discussion. We do so by focusing on key phenomena involving international marketing and international management more broadly that might be affected both by spatial proximity and by location-specific influences and for which data are publicly available. We consider three phenomena and corresponding independent variables. For all three analyses the chief independent variable is the measure of geographic distance that we also analysed for Table 1. In addition, all three analyses follow the guideline of considering minimum seven 'origin' and seven 'destination' countries or 49 country dyads (Van Hoorn and Maseland, 2016; see, also, Franke and Richey, 2010). In fact, two out of three samples that we study also meet Van Hoorn and Maseland's (2016) stricter guideline of considering minimum 10 'origin' and 10 'destination' countries or 100 country dyads.

3.2.1 Dependent variables, samples, and empirical approach

3.2.1.1 Dependent variables

The empirical illustration focuses on two main firm-level dependent variables involving the importance of different sources of information on new customers for foreign subsidiaries. The first of these variables concerns the importance of business associations as potential sources of information for a foreign subsidiary and the second concerns the importance of trade fairs. We consider these two variables because they speak to how multinational establishments go about marketing their products and engaging with customers in their host countries. Finally, to be complete, we also consider inward FDI as the dependent variable. The reasons are that FDI is an important entry mode (Budeva and Torres-Baumgarten, 2021) and that FDI was also studied in Van Hoorn's (2020) critique of empirical cultural and institutional distance research.

3.2.1.2 Samples and data sources[3]

Data for the two firm-level dependent variables come from the 2005 wave of the Business Environment and Enterprise Performance Survey or BEEPS (EBRD, 2005), which is increasingly used in international marketing and management (Abumousa, 2018; Bertomeu, 2018). The BEEPS is organised by the World Bank and the European Bank for Reconstruction and Development (EBRD). It asked official representatives of establishments in more than 25 (transition) economies various questions about their environment for doing business and their establishment's performance. We select the 2005 wave because it is the most recent wave that included a survey item on foreign ownership of the establishment and also an item on the nationality of the foreign owner. As in Spencer and Gomez (2011), combining these two items enables us to identify home and host countries for (partially) foreign-owned establishments in the BEEPS dataset. The two dependents derive from the generic item asking respondents to rate the importance of different potential sources of information on new customers for their firm on a discrete scale that ranges from 1, 'Not important' to 5, 'Extremely important.' The item covers different potential sources, not least business associations and trade fairs as studied by us. Following standard definitions of multinational companies and FDI, we only consider establishments with minimum 10% foreign ownership. This renders samples of minimum 766 multinational establishments, comprising minimum 39 home and 28 host countries and at least 281 home-host country dyads.

Data for inward FDI come from the OECD (2020) and refer to the percentage of a particular country's total FDI stock that is in a particular destination country (among a group of eight possible destination countries). In contrast to Van Hoorn (2020), we consider FDI data for the year 2017 instead of 2018. The reason is that this renders a larger sample, comprising 279 observations or dyads of FDI coming from 36 possible 'origin' countries and located in eight possible 'destination' countries.

Since our interest is in potential statistical biases in the estimated relationships, we do not formulate hypotheses about the likely effect of geographic distance on the three dependent variables that we consider. However, following prior work on distance as a challenge to (multinational) firms (Zaheer, 1995), we expect that geographic distance makes doing business abroad more difficult. Hence, we expect that as geographic distance increases, the importance of formal sources of customer information such as business associations or trade fairs increases as well, particularly compared to the importance of sources of information that are less formal such as family, friends or former employees. The underlying logic is that business association and trade fairs substitute for these other sources of information, which are more likely to reside in the firm's home country and are likely less able to provide information on customers in a specific host country when their spatial proximity to this host country decreases. Similarly, we expect that the attractiveness of a host country decreases as geographic distance increases, meaning that geographic distance reduces inward FDI.

3.2.1.3 Empirical approach

To set a benchmark for judging possible biases caused by location-identification, we apply the two empirical approaches to overcoming the location-identification problem proposed by Van Hoorn (2020). The first approach is called the fixed-effects approach and involves inclusion of location fixed effects, i.e., dummies for destination and origin locations, as control variables. The rationale of this approach is that these added control variables capture any location-specific effects that would otherwise have been captured by spatial proximity (e.g., country-specific measurement error), thus rendering an unbiased estimate of effects genuinely due to spatial proximity.[4] Van Hoorn's (2020) second proposed approach involves creating a 'pure' proximity measure, meaning a proximity measure that is cleansed from location-specific influences. Implementation of this 'pure-distance' approach involves two steps. The first step is to estimate an OLS model with spatial proximity as the dependent variable and origin and destination location fixed effects as the independent variables and save the residuals. These residuals provide the corrected measure of proximity. The second step is to use this pure, corrected measure of spatial proximity instead of the raw, uncorrected measure. [5] By construction, using such a pure spatial proximity measure renders the same estimated coefficient for spatial proximity as obtained when adding location fixed effects as control variables (Van Hoorn, 2020). However, the use of a pure proximity measure retains the possibility of empirically considering substantive location-specific influences that are also of interest to many researchers and practitioners in the field, which is not possible when adding location fixed effects.[6] Meanwhile, comparing results obtained using these two approaches to results obtained using an ordinary empirical model shows us the extent to which location-identification leads to biased estimates for the effect of spatial proximity on the selected dependent variables.

3.2.2 Results

Results confirm that location-identification causes a bias in the estimated relationship between geographic distance and different phenomena in international marketing and management. In case of the importance of business associations as a source of information on new customers, the baseline model that does not control for location-specific influences suggests a strong, statistically highly significant positive effect of geographic distance on the importance of business associations (Table 2, Model 1). Estimated this way, geographic distance seems to account for some 1.9% of total variation in the importance of business associations. However, this estimate includes the confounding influence of the location-specific effects captured by the uncorrected measure of geographic distance. Getting rid of these confounding influences, either by including location dummies (Model 3) or by using a pure proximity measure (Model 4), variance explained by geographic distance decreases substantially to about 0.09% (bottom row of Table 2). In addition, although the estimated coefficient for geographic distance remains positive, it is no longer statistically significant at usual levels ($p > 0.1$ instead of $p = 0.0000$).

Table 2 Illustration of possible statistical biases when using spatial proximity as an independent variable: evidence from the importance of business associations

Dependent = Importance of business associations as potential source of information on new customers	*Model 1*	*Model 2*	*Model 3*	*Model 4*	*Model 5 (=Model 4 bootstrapped with 10,000 repetitions)*
Geographic distance	0.1393 [p = 0.0000]		0.1484 [p = 0.2965]		
Pure, corrected geographic distance	-	-	-	0.1484 [p = 0.3815]	0.1484 [p = 0.3823]
Location fixed effects included	No	Yes	Yes	No	No
No. of observations	918	918	918	918	918
No. of dyads	315	315	315	315	315
No. of host locations	28	28	28	28	28
No. of home locations	40	40	40	40	40
R^2	1.941%	12.59%	12.68%	0.091%	0.091%
ΔR^2 due to geographic distance	1.941%	-	0.091%	0.091%	0.091%

Notes: Table reports estimated coefficients (and robust p-values) for an OLS regression. The estimated models vary in two possible ways: 1) the independent variables included, only spatial proximity (Model 1), only country fixed effects (Model 2) or both spatial proximity and country fixed effects (Model 3); 2) the specific measure of spatial proximity used as an independent variable, pure corrected distance (Models 4 and 5) or raw geographic distance (Models 1 and 3). The pure, corrected measure of geographic distance refers to the residuals of a simple OLS regression with geographic distance as the continuous dependent variable and location fixed effects as the independent variables (see main text and Table S1 in the online supplement). Model 5 repeats Model 4 but uses bootstrapping to preempt possible biases caused by the fact that the pure indicator of geographic distance is constructed using regression analysis (see Note 5).

Considering the importance of trade fairs as the dependent variable, we find a similar upward bias in the estimated effect of geographic distance Table 3. The naïve estimate (Model 6) indicates a reasonably strong relationship between geographic distance and the importance of trade fairs that is statically significant at usual levels ($p < 0.05$). As before, however, controlling for location-identification substantially reduces the amount of variance that is explained by geographic distance and renders estimates that are not statistically significant at usual levels (Models 8 and 9). In general, even though the estimated coefficients for geographic distance do not change much in Tables 2 and 3, it seems that in these cases the estimated effects of raw, uncorrected geographic distance are picking up effects due to country location. Implication is that a seemingly reliably estimated effect of geographic distance is not a genuine distance effect but a location effect.

Table 3 Illustration of possible statistical biases when using spatial proximity as an independent variable: evidence from the importance of trade fairs

Dependent = Importance of trade fairs as potential source of information on new customers	*Model 6*	*Model 7*	*Model 8*	*Model 9*	*Model 10 (=Model 9 bootstrapped with 10,000 repetitions)*
Geographic distance	0.0683 [p = 0.0318]		0.0811 [p = 0.5655]		
Pure, corrected geographic distance				0.0811 [p = 0.6353]	0.0811 [p = 0.6353]
Location fixed effects included	No	Yes	Yes	No	No
No. of observations	914	914	914	914	914
No. of dyads	313	313	313	313	313
No. of host locations	28	28	28	28	28
No. of home locations	40	40	40	40	40
R^2	0.467%	13.39%	13.42%	0.028%	0.028%
ΔR^2 due to geographic distance	0.467%	-	0.028%	0.028%	0.028%

Note: Table reports estimated coefficients (and robust p-values) for an OLS regression. The estimated models vary in two possible ways: 1) the independent variables included, only spatial proximity (Model 6), only country fixed effects (Model 7) or both spatial proximity and country fixed effects (Model 8); 2) the specific measure of spatial proximity used as an independent variable, pure corrected distance (Models 9 and 10) or raw geographic distance (Models 6 and 8). The pure, corrected measure of geographic distance refers to the residuals of a simple OLS regression with geographic distance as the continuous dependent variable and location fixed effects as the independent variables (see main text and Table S2 in the online supplement). Model 10 repeats Model 9 but uses bootstrapping to preempt possible biases caused by the fact that the pure indicator of geographic distance is constructed using regression analysis (see Note 5).

Interestingly, results indicate that location-identification does not cause an upward but a downward bias in the estimated effect of geographic distance on inward FDI stocks Table 4. In our sample, geographic distance has a negative effect on inward FDI (Model 11) and this effect becomes more strongly negative when taking into account that geographic distance correlates with location and partly captures location-specific effects

(Models 13 and 14). According to the naïve, baseline estimate, geographic distance accounts for some 2.7% of total variation in inward FDI (Model 11). However, this estimate includes the confounding influence of the location-specific effects captured by the uncorrected measure of geographic distance. Getting rid of these confounding influences, variance explained by geographic distance increases more than twofold to about 6.3% (Models 13 and 14). Hence, it seems that in this case the estimated relationship between raw, uncorrected geographic distance and FDI is capturing two opposing influences that have a net effect on inward FDI that is also negative but smaller than the 'true' negative effect of geographic distance on inward FDI.

Table 4 Illustration of possible statistical biases when using spatial proximity as an independent variable: evidence from inward FDI

Dependent = Inward FDI stock	*Model 11*	*Model 12*	*Model 13*	*Model 14*	*Model 15 (=Model 14 bootstrapped with 10,000 repetitions)*
Geographic distance	–0.1667		–0.4381		
	[p = 0.0028]		[p = 0.0031]		
Pure, corrected geographic distance	–	–	-	–0.4381	–0.4381
				[p = 0.0087]	[p = 0.0080]
Location fixed effects included	No	Yes	Yes	No	No
No. of observations/dyads	279	279	279	279	279
No. of destination locations	8	8	8	8	8
No. of origin locations	36	36	36	36	36
R^2	2.719%	38.12%	44.42%	6.298%	6.298%
ΔR^2 due to geographic distance	2.719%	–	6.298%	6.298%	6.298%

Notes: Table reports estimated coefficients (and robust p-values) for an OLS regression. The pure, corrected measure of geographic distance refers to the residuals of a simple OLS regression with geographic distance as the continuous dependent variable and location fixed effects as the independent variables (see main text and Table S3 in the online supplement). Model 15 repeats Model 14 but uses bootstrapping to preempt possible biases caused by the fact that the pure indicator of geographic distance is constructed using regression analysis (see Note 5).

3.2.3 Discussion of the empirical results on location-identification and (Un)biased estimates

Of course, the above results do not prove that every analysis using spatial proximity as an independent variable is biased. Together, however, these results provide a clear illustration of how location-identification can introduce important upward or downward biases in estimated effects of spatial proximity. Hence, the conclusion that using spatial proximity as an independent variable is not as innocuous as it may seem.

We should also note, though, that the bias caused by location-identification may partly be addressed by controlling for observable location factors. Country-level examples of such observable location factors include measures of gross domestic product (GDP) per capita, investment regulations, institutional quality, et cetera. When it comes to empirical model specification, researchers typically have good theoretical reason to assume that certain location characteristics affect the dependent variable of interest and may bias the estimated distance effect if they are not controlled for.[7] Unfortunately, however, inclusion of observable confounders such as per-capita GDP is unlikely to get rid of biases caused by unmeasurable and unobserved location-specific confounders captured by distance, not least of which is random, location-specific measurement error. In fact, the power of including location fixed effects is precisely that they capture the influence of location-specific factors that are inherently unmeasurable. Hence, even when there is grounds for thinking that a specific set of observable country-level control variables will eliminate all biases due to location-identification, we would still need the fixed-effects or pure-distance approaches to check whether this is indeed the case. Specifically, these two approaches provide the benchmark of unbiased estimates against which estimates obtained using models that only control for observable country-level confounders need to be validated.

4 Discussion

Spatial proximity has been linked to a wide range of important phenomena, not just in international marketing and management but also in other fields of business and management research. However, drawing on recent insights from the literature on contextual distance in international management (Brouthers et al., 2016; Van Hoorn, 2020), this paper finds that the use of spatial proximity as an independent variable in empirical analyses is subject to a little recognised but critical challenge. This challenge is that, in real-world samples, variation in measured spatial proximity often comprises significant amounts of variation that is specific to the locations in the sample. Location-specific variation in turn means that spatial proximity correlates with location, thus partly capturing location-specific effects and causing biased model estimates when used as an independent variable. This confounding effect is particularly relevant for export managers or managers of MNEs. Properly managing firms' international endeavors requires being able to identify whether challenges that firms face are driven by geographic distance or by location-specific factors.

Two main objections to this paper are that it lacks academic relevance and the potential for multicollinearity when adding country or location fixed effects as control variables. Concerning the issue of relevance the objection would be twofold. On the one hand it would be that the international marketing and international management literature does not have much interest in spatial proximity, preferring analyses of cultural and institutional distance instead. On the other hand, it would be that present-day distance studies in (international) management already rely on panel data and control for country or location fixed effects as a matter of course, rendering the location-identification problem mute. In response, we note that new papers focusing on spatial proximity are still being published and that many (international) marketing and management studies do consider spatial proximity (even if only as a control variable). Similarly, we note that whereas some recent distance studies include country fixed effects, panel data remain scarce and this practice is still far from standard. In addition, we note that spatial proximity is considered in many fields of research, not just international marketing or international management. Finally, we note that spatial proximity's ability to capture location-specific effects does not only have implications for the design of future studies but also affects our understanding of existing studies. In particular, we need to recognise that prior studies may lack location-level control variables needed to ensure that empirical results found are not biased or even spurious. Hence, some prior distance studies are in fact less valid empirically than they may have seemed in the past.

Concerning the issue of multicollinearity we would point particularly to Lindner et al.'s (2020) recent work, which dispels various myths about multicollinearity in (international) management research. Most importantly, multicollinearity does not involve a statistical bias of any sort. Instead, multicollinearity involves imprecise estimates (i.e., high standard errors and thus high p-values) that are in turn caused by a lack of variation in the sample. In short, collinearity between variables means that it becomes more difficult to estimate the effect of one variable independent from the effect of another, correlated variable. The essential solution to this problem is the collection of additional data, which is why several standard statistics textbooks prefer speaking of so-called micronumerosity rather than of multicollinearity (see, e.g., Wooldridge, 2002). Excluding control variables that are collinear with the independent variable of interest, in contrast, is not a viable solution. The reason is simply that removing (a) variable(s) from a properly specified empirical model causes a bias in the estimated coefficients for this model (Lindner et al., 2020; Wooldridge, 2002).

5 Recommendation and concluding remarks

Based on the findings from this study, our generic recommendation is that spatial proximity studies consider the possibility of location-identification and the resulting endogeneity problem as a matter of course. Practically, a first step is that researchers consider the extent to which variation in spatial proximity in their sample is location-specific and correlates with origin or destination location. As illustrated by the evidence presented in this paper, location-identification is not uncommon. Hence, testing for location-specific variation or estimating multinomial models with location as the dependent variable provides a means to identify a potential threat to the validity of estimated effects of spatial proximity found later on. If results of Step 1 give cause for concern, Step 2 involves taking action to reduce statistical biases due to location-identification and thus improve the validity of found effects of spatial proximity. The generic solution to the endogeneity problem that may be caused by location-identification is finding an appropriate way of dealing with correlated measurement error. Statistics textbooks discuss several standard solutions (see, for example, Wooldridge, 2002). However, in case of spatial proximity and location-identification, we know exactly what the measurement error is namely that measured proximity correlates with location. Hence, we recommend using one of the standard approaches in the literature (include dummy variables to control for location fixed effects [the fixed-effects approach] or use pure spatial proximity measures that are cleansed from confounding location effects [the pure-distance approach]).

Overall, we are far from pessimistic about the future for spatial proximity studies in (international) management research. Location-identification may seem challenging but different solutions to this challenge exist and these can be implemented relatively straightforwardly. At the same time, we call for further research on location-identification and possible conflation of distance effects with location effects. This paper has focused on a specific type of spatial proximity studies involving clearly defined home and host locations and associated dyads. This type of spatial proximity studies is dominant in the literature. However, there is also an increasing number of studies that does not consider distance per se but so-called 'added distance' (e.g., Hendriks, 2020; Hutzschenreuter et al., 2014; Schu and Morschett, 2017). Distance as considered in this paper revolves around origin-destination location dyads and the distance between the origin and the destination location that together form a dyad. Added distance, in contrast, revolves around firms and their portfolio of cross-border value-added activities. Added distance can thereby be defined as the distance that is added when a firm expands its foreign activities beyond its most distant existing foreign activity. Added distance is thus not defined by two locations in the same way that distance is. This difference, in turn, could mean that empirical added distance studies are less vulnerable to conflating effects of distance, specifically of added distance, with effects due to the specific foreign locations in which a firm is active. Hence, an interesting avenue for future research is to assess how controlling for (unmeasurable) location-specific factors affects estimates of the effect of added distance on various phenomena of interest. We further find that all the work so far has focused on objective distance indicators and has disregarded subjective or perception-based distance indicators. Psychic distance, however, is a quintessential construct in international marketing and management (Szylit and Botelho, 2017). Hence, we think that is also very worthwhile to check for the occurrence and extent of statistical biases in empirical studies that use this construct.

References

Abumousa, A. (2018) 'Firm innovation in the Asia and Pacific region: the role of governance environment, firm characteristics, and external finance', *International Journal of Export Marketing*, Vol. 2, No. 3, pp.180–209.

Becker, T.E. (2005) 'Potential problems in the statistical control of variables in organizational research: a qualitative analysis with recommendations', *Organizational Research Methods*, Vol. 8, No. 3, pp.274–289.

Bertoli, S., Goujon, M. and Santoni, O. (2016) *The CERDI-Seadistance Database* [online] https://ferdi.fr/en/indicators/the-cerdi-seadistance-database (accessed 17 April 2020).

Bertomeu, G.S.M. (2018) 'Analysis of the export performance of state-owned enterprises from Latin American and Caribbean, reviewing corruption and informal economy', *International Journal of Export Marketing*, Vol. 2, No. 4, pp.264–290.

Beugelsdijk, S., Kostova, T. and Roth, K. (2017) 'An overview of Hofstede-inspired country-level culture research in international business since 2006', *Journal of International Business Studies*, Vol. 48, No. 1, pp.30–47.

Brouthers, L.E., Marshall, V.B. and Keig, D.L. (2016) 'Solving the single country sample problem in cultural distance studies', *Journal of International Business Studies*, Vol. 47, No. 4, pp.471–479.

Budeva, D. and Mullen, M.R. (2016) 'Does culture matter for international market selection?', *International Journal of Export Marketing*, Vol. 1, No.. 2, pp.193–214.

Budeva, D. and Torres-Baumgarten, G. (2021) 'The effect of institutional distance on international market selection: comparing export to foreign direct investment', *International Journal of Export Marketing*, Vol. 4, No. 2, pp.127–149.

Cannon, J.P. and Homburg, C. (2001) 'Buyer–supplier relationships and customer firm costs', *Journal of Marketing*, Vol. 65, No. 1, pp.29–43.

Cramer, D. and Howitt, D.L. (2004) *The Sage Dictionary of Statistics: a Practical Resource for Students in the Social Sciences*, Sage, London.

EBRD (2005) *Business Environment and Enterprise Performance Survey (BEEPS)* [onlinehttps://www.beeps-ebrd.com/data/2005.

Everitt, B. (2002) *The Cambridge Dictionary of Statistics*, 2nd ed., Cambridge University Press, Cambridge.

Feenstra, R. (2004) *Advanced International Trade: Theory and Evidence*, Princeton University Press, Princeton.

Franke, G.R. and Richey, R.G. (2010) 'Improving generalizations from multi-country comparisons in international business research', *Journal of International Business Studies*, Vol. 41, No. 8, pp.1275–1293.

Funk, R.J. (2014) 'Making the most of where you are: geography, networks, and innovation in organizations', *Academy of Management Journal*, Vol. 57, No. 1, pp.193–222.

Ghemawat, P. (2016) *The Laws of Globalization and Business Applications*, Cambridge University Press, Cambridge.

Head, K. and Mayer, T. (2014) 'Gravity equations: workhorse, toolkit, and cookbook', In *Handbook of International Economics*, Vol. 4, pp.131–195, Elsevier, Amsterdam.

Hendriks, G. (2020) 'How the spatial dispersion and size of country networks shape the geographic distance that firms add during international expansion', *International Business Review*, Vol. 129, No. 6, p.101738.

Hutzschenreuter, T., Kleindienst, I. and Lange, S. (2014) 'Added psychic distance stimuli and MNE performance: Performance effects of added cultural, governance, geographic, and economic distance in MNEs' international expansion', *Journal of International Management*, Vol. 20, No. 1, pp.38–54.

Knoben, J. and Oerlemans, L.A. (2006) 'Proximity and inter-organizational collaboration: a literature review', *International Journal of Management Reviews*, Vol. 8, No. 2, pp.71–89.

Lindner, T., Puck, J. and Verbeke, A. (2020) 'Misconceptions about multicollinearity in international business research: identification, consequences, and remedies', *Journal of International Business Studies*, Vol. 51, No. 3, pp.283–298.

Mayer, T. and Zignago, S. (2011) *Notes on CEPII's Distances Measures: The GeoDist Database* [online] https://scholar.google.nl/scholar?cites=5585266570934732236&as_sdt=2005&sciodt=0,5&hl=nl.

OECD (2020) *Inward FDI Stocks By Partner Country (Indicator),* DOI: 10.1787/a1818a82-en.

Ragozzino, R. and Reuer, J.J. (2011) 'Geographic distance and corporate acquisitions: signals from IPO firms', *Strategic Management Journal*, Vol. 32, No. 8, pp.876–894.

Schu, M. and Morschett, D. (2017) 'Foreign market selection of online retailers-a path-dependent perspective on influence factors', *International Business Review*, Vol. 26, No. 4, pp.710–723.

Spencer, J. and Gomez, C. (2011) 'MNEs and corruption: the impact of national institutions and subsidiary strategy', *Strategic Management Journal*, Vol. 32, No. 3, pp.280–300.

Szylit, F. and Botelho, D. (2017) 'An experiment on the effect of psychic distance on internationalisation of retailers', *International Journal of Export Marketing*, Vol. 1, No. 4, pp.396–414.

Todri, V., Adamopoulos, P. and Andrews, M. (2022) 'Is Distance really dead in the online world? The moderating role of geographical distance on the effectiveness of electronic word of mouth', *Journal of Marketing*, Vol. 86, No. 4, pp.118–140.

Van Hoorn, A. (2020) 'Cross-national distance as an explanatory variable in international management', *Journal of International Management*, Vol. 26, No. 3, p.100773.

Van Hoorn, A. and Maseland, R. (2014) 'Is distance the same across cultures? A measurement-equivalence perspective on the cultural distance paradox', *Multinational Enterprises, Markets and Institutional Diversity*, Vol. 9, No. 1, pp.207–227.

Van Hoorn, A. and Maseland, R. (2016) 'How institutions matter for international business: Institutional distance effects vs. institutional profile effects', *Journal of International Business Studies*, Vol. 47, No. 2, pp.374–381.

Wooldridge, J.M. (2002) *Econometric Analysis of Cross Section and Panel Data,* MIT Press, Cambridge.

Zaheer, S. (1995) 'Overcoming the liability of foreignness', *Academy of Management Journal*, Vol. 38, No. 2, pp.341–363.

Notes

1 Other intuitive operationalisations of the theoretical construct of spatial proximity include such measures as road distance, air distance, or the travel time between two locations. When considering the concept of distance or proximity, international marketing and management researchers also often consider non-geographical types of distance such as cultural distance.

2 Figure S1 in the online supplement illustrates how distance in two-dimensional (geographical) space can have infinite directions whereas difference scores or distance in one-dimensional space can have only two directions. Figure S2 illustrates how large distances in one-dimensional space that start from a location close to the extreme of the distribution can only go in one direction, thus providing a close-to-exact identification of the origin and destination locations involved.

3 The online supplement presents details and descriptive statistics for the three samples (Tables S1–S3 in Sections S2–S4).

4 In terms of equations (1a) and (1b), adding location dummies or fixed effects overcomes the location-identification problem because it changes the regression model in equation (1a) in such a way that it has a new error term, e_N, that, unlike the original error term (e), no longer correlates with the independent variable I: $D = \beta_0 + \beta_1 I + L + e_N$ and corr(I, e_N) = 0 (where L denotes location fixed effects).

5 Preempting possible biases, it could be useful to use standard bootstrapping techniques when estimating empirical models that include a corrected measure of spatial proximity. In terms of Eqs. 1a and 1b, the pure-distance approach overcomes the location-identification problem because it results in a new independent variable, IN, that, unlike the original independent variable (I), is no longer correlated with the error term (e) of the regression model in equation equation (1a) $D = \beta_0 + \beta_1 I_N + e$ and corr(I_N, e) = 0.

6 Of course, longitudinal data would make it possible to estimate models that control for origin and destination location fixed effects while also considering the influence of time-varying location-specific factors. In fact, this is textbook practice in the economics literature estimating so-called gravity equations, for instance, for longitudinal data on bilateral trade flows (Feenstra, 2004). Countries differ from each other in a variety of dimensions and quite a few of these dimensions may affect economic interactions between these countries. Controlling for location fixed effects helps make sure that gravity equations capture genuine effects of geographical distance and not the influence of such factors as the population size of a country or the size of its national income (e.g., Head and Mayer, 2014). Longitudinal data, however, are much rarer in international marketing and management research than cross-sectional data are (see Budeva and Mullen, 2016 for an interesting exception).

7 To be sure, inclusion of, say, host-country per-capita GDP only affects the estimated coefficient for distance if host-country per-capita GDP correlates both with distance and with the dependent variable (see, also, the discussion on correlated measurement error in Section 3.1). On the other hand, if host-country per-capita GDP solely correlates with the dependent variable, there is no statistical reason for including it as a control variable. As discussed extensively in the methods literature (see, e.g., Becker, 2005), in this latter scenario,

inclusion of host-country per-capita GDP will improve model fit but does not reduce statistical biases nor affect the estimated distance coefficient. At the same time, controlling for host-location observables of course does not help addressing biases caused by home-location observables that correlate with both distance and the dependent variable. Distance studies in international marketing and management often do not control for both home-location and host-location observables simultaneously, however.